# Proceedings of the GIS Research UK 18th Annual Conference GISRUK 2010

University College London
14th – 16th April 2010

Editors: Mordechai (Muki) Haklay
Jeremy Morley
Hanif Rahemtulla

**Proceedings of the GIS Research UK**
**18th Annual Conference**
**GISRUK 2010**

University College London
14th – 16th April 2010



Published April 2010, University College London

# Welcome

The GIS Research UK conference is now in its 18th year, and for the second time in its history, in London. As in previous years, the conference is going from strength to strength with a wide range of papers and topics. As the field of research into the science and applications of geographical information systems continue to evolves, so does GISRUK. This year, the papers are covering a wide range of subjects from modelling and simulation to the emerging GeoWeb.

GISRUK is a special conference. While originally created as the UK's national GIS research conference, and established in 1993, it became recognised as one of the leading research conference in the area of GIScience worldwide. They attract delegates from all parts of the UK, together with the European Union and further afield and from disciplines including Geography, Computer Science, Planning, Archaeology, Geology, Geomatics and Engineering. They are organised and managed by a team of volunteers and the strength of GIS research in the UK is demonstrated by the fact that for the past 18 conference, it was not hosted in the same institution twice (and only twice now in the same city – London).

The GISRUK conferences have the following aims:

- to act as a focus for GIS research in the UK
- to provide a mechanism for the announcement and publication of GIS research
- to act as an interdisciplinary forum for the discussion of research ideas
- to promote active collaboration amongst researchers from diverse parent disciplines
- to provide a framework in which postgraduate students can see their work in a national context

All the papers that have been submitted to the conference have been reviewed by at least two reviewers, with the vast majority receiving three reviews. The papers in this collection have been revised in the light of reviewers' comments. This is an important aspect of the communal learning process that GISRUK fosters. We would like to thank the 150 reviewers who voluntarily helped in evaluating the papers.

The organisation of the conference would not be possible without the support of the sponsors. We are therefore pleased to acknowledge the kind sponsorship that enables the conference to make it affordable to young researchers and academics, invite leading keynote speakers and provide awards. This year, we have received support from the Ordnance Survey, the Association for Geographic Information, ESRI (UK), EDINA, Cadcorp, the Geoinformation Group, Taylor & Francis, Wiley, MIMAS and GEES.

Welcome to London and UCL and we hope that you enjoy the conference!

*Muki Haklay*
*Co-Chair, Local Organizing Committee*
*Department of Civil, Environmental and Geomatic Engineering, UCL*
*Jeremy Morley*
*Co-Chair, Local Organizing Committee*
*University of Nottingham*

## GISRUK National Organising Committee

| | |
|---|---|
| Katherine Arrell | University of Leeds |
| Angela Baker | Association for Geographic Information |
| Jane Drummond | University of Glasgow |
| Christine Dunn | University of Durham |
| David Fairbairn | University of Newcastle |
| Bruce Gittings (Chair) | University of Edinburgh |
| Peter Halls | University of York |
| Glen Hart | Ordnance Survey |
| Muki Haklay | University College London |
| David Lambrick | Manchester Metropolitan University |
| Jeremy Morley | University of Nottingham |
| Nick Mount | University of Nottingham |
| Nick Tate | University of Leicester |
| Adam Winstanley | National University of Ireland, Maynooth |
| Steve Wise | University of Sheffield |
| Jo Wood | City University |

## GISRUK Local Organising Committee

| | |
|---|---|
| Spencer Chainey | Jill Dando Institute of Crime Science |
| Tao Cheng | Civil, Environmental and Geomatic Engineering |
| Claire Ellul | Civil, Environmental and Geomatic Engineering |
| Muki Haklay (co-chair) | Civil, Environmental and Geomatic Engineering |
| Shane Johnson | Jill Dando Institute of Crime Science |
| Liz Jones | Civil, Environmental and Geomatic Engineering |
| Paul Longley | Geography / CASA |
| Pablo Mateos | Geography |
| Jeremy Morley (co-chair) | University of Nottingham |
| Hanif Rahemtulla | Civil, Environmental and Geomatic Engineering |
| Alex Singleton | Geography / CASA |
| Mike de Smith | Civil, Environmental and Geomatic Engineering /CASA |
| Helena Titheridge | Civil, Environmental and Geomatic Engineering |

# Contents

## Session 1A *London as a Global City*



## Session 1B *Visualisation in Modelling and Simulation*



## Session 2A *Crime and Place*

## Session 2B ***Pedestrian and Urban Modelling***



## Session 3A ***Health Geography***



## Session 3B ***Volunteered Geographic Information I***

## Session 4A ***Geodemographics***



## Session 4B ***Volunteered Geographic Information II***



## Session 5A ***Environmental GIS***

## Session 5B ***GeoWeb I***



## *Session 6A* ***Short Papers***



## Session 7A ***Transport***

## Session 7B ***Geoweb II***



## Session 8A ***Urban Topology and Morphology***



## Session 8B ***HCI and GIS I***



## Session 9A ***Automatic Mapping***

## Session 9B ***HCI and GIS II***



## ***Posters' Abstracts***

# Location Intelligence: an innovative approach to business location decision making

Patrick Weber[1], Dave Chapman[2]

[1]Department of Computer Science, University College London, Gower Street London, WC1E 6BT
Tel: +44 (0)20 7718 5430 | Email: p.weber@ucl.ac.uk | Fax: +44 (0) 20 7813 2843

[2]Department of Management Science and Innovation, University College London, Gower Street, London, WC1E 6BT
Tel: +44 (0)20 7679 0441 | Email: d.chapman@ucl.ac.uk | Fax: +44 (0)20 7679 3209

KEYWORDS: Inward Investment, Business GIS, SDSS, Regional Development, Multi-Criteria Decision Making

## 1. Introduction

As one of the leading 'world cities' London is home to a highly internationalised workforce and is particularly reliant on these sources of foreign direct investment (FDI). In the face of increasing global competition and a very difficult economic climate, the capital must compete effectively to encourage and support such investors.

Through a collaborative study with London's official foreign direct investment agency, Think London, the need for a coherent framework for data, methodologies and tools to inform business location decision making became apparent (Weber & Chapman 2009). In this paper we discuss the development of a rich environment to iteratively explore, compare and rank London's business neighbourhoods. This is achieved through the development, integration and evaluation of data and its manipulation to form a model for locational based decision support.

First, we discuss the development of a geo-business classification for London which draws upon methods and practices common to many geospatial neighbourhood classifications that are used for profiling consumers. In this instance a geo-business classification is developed by encapsulating relevant location variables using Principal Component Analysis into a set of composite area characteristics. Second, we discuss the implementation an appropriate Multi-Criteria Decision Making methodology, in this case Analytical Hierarchy Process (AHP), enabling the aggregation of the geo-business classification and decision makers preferences into discrete decision alternatives (Carver 1991; Jankowski 1995). Lastly, we present the preliminary results of the integration of both data and model through the development and evaluation of a web-based prototype and evaluate its usefulness through scenario testing.

## 2. A geo-business classification for London

Following on from previous work by Weber and Chapman (2009), a database of 50 variables (see Table 1) was developed that supports four major domains of locational knowledge needed for FDI promotion activities in Greater London:

1. the discovery, quantification and qualification of industry sector clusters (*Companies*),
2. the characterisation of the available talent pool and daytime population (*Working Population*),
3. quantity and quality of the *Property Stock*,
4. a more general appreciation of the *Living Environment* of London neighbourhoods.

These variables, recorded at disparate geographical levels, necessitated a coherent geographical framework able to summarise the historic urban development of Greater London as many individual towns and cities (Hebbert 1998; Thurstain-Goodwin & Unwin 2000; Ackroyd 2001; URBED 2002) with persisting separate identities and differing characteristics, in an economic sense, as well as socially, demographically and culturally.

**Table 1.** Initial Spatial Database

| Class | Source | Variables | | Geography |
|---|---|---|---|---|
| Companies | Annual Business Inquiry 2007 | Creative industries | Environmental | LSOA |
| | | Higher Education & Research | Construction | |
| | | Health | Retail | |
| | | Social work | Transport & logistics | |
| | | Tourism & leisure | Charity & voluntary | |
| | | Utilities | Life sciences | |
| | | Professional services | Pharmaceuticals | |
| | | Financial services | Medical equipment | |
| | | Food & drink | Manufacturing | |
| | | ICT | Real estate | |
| | | Ratio of Workplaces over Employees | | |
| Working Population | Census 2001: Special Workplace Statistics | Higher managerial and professional Occupations: Large employers and higher managerial Occupations Higher professional Occupations | | Census Wards |
| | | Lower managerial and professional Occupations | | |
| | | Intermediate Occupations | | |
| | | Small employers and own account workers | | |
| | | Lower supervisory and technical occupations | | |
| | | Semi-routine occupations | | |
| | | Routine Occupations | | |
| | | Never worked and long-term unemployed | | |
| Property Stock | Rateable Value Statistics 2007 | Rateable Value per square meter - Offices | | MSOA |
| | | Rateable Value per square meter - Premises | | |
| | | Rateable Value per square meter - Factories | | |
| | | Rateable Value per square meter - Warehouses | | |
| | | Total Floorspace - Offices | | |
| | | Total Floorspace - Retail Premises | | |
| | | Total Floorspace - Factories | | |
| | | Total Floorspace – Warehouses | | |
| Living Quality | Index of Multiple Deprivation 2007 | Overall Score | | LSOA |

The Town Centre boundaries dataset (Thurstain et al. 2001; Lloyd 2004), comprising 206 Greater London Town Centres, was chosen as the most appropriate quantitative expression of the urban nucleus. Aggregating the database to this common geography allowed us to develop quantifiable and comparable indicators, fit to constitute the basis of geo-business area classification for Greater London.

Principal Components Analysis was then used as an investigative methodology to identify patterns of variance among location variables, enabling the simplification and generalisation of the town centre database. Nine principal components were selected which retained 64 percent of the original variance. These components describe distinctive domains of business activities, allowing the meaningful characterisation and comparison between FDI location options at the scale of the town centre.

Each geo-business class is characterised by a memorable and distinctive class name, some street views of typical Town Centre activities along with a narrative highlighting the most distinctive characteristics in terms of of economic activity, liveability, socio-economic makeup of the workforce and property stock. The nine component descriptions are:

- **Urban Professionals:** Mainly professional and financial service type companies, representing a mix of large & small employers.The workforce is composed of skilled managerial and professional employees. The urban environment is characterised by high quality offices, with only limited retail space available.
- **Blue Collar Industry:** Predominantly manufacturing, food and drink type activities are concentrated here, along with logistics and distribution. There is a mix of large and small employers, which employ mostly lower skilled routine and technically skilled people. Mainly warehousing, with only limited office space available.
- **Blue Chip Finance:** Significant concentration of large financial services companies, which employ managerial and professionally skilled workforce. These areas are characterised by mostly office buildings and not a lot of tourism activities and a lack of small and independent companies or self employed workers.
- **Third Sector Centres:** Characterised mostly by the presence of social services, charity, voluntary and other third sector type organisations. Located in a deprived environment, with low value/quality retail and office premises.
- **Big Sheds and Trucks:** Mostly logistics and distribution type companies which employ lower skilled workers. The available commercial spaces are mainly warehouses and factories, of good quality, but only limited office space available. There are almost no retail activities or financial services type companies.
- **High End Streets:** High streets containing high value shops, retail related activities as well as estate agents. Also attractive to tourists, evidenced by significant tourism activities. The workforce is composed mainly of professionals, working in high value retail space or offices.
- **Creative and Green Minds:** Environment with a concentration of creative industry, ICT and environmental industry type businesses. A noted absence of healthcare related activities. Mostly larger employers which employ a highly qualified workforce, with little manual or lowly skilled labour.
- **Sights of London:** Areas focused around tourism and retail related activities. Also a noted concentration of high quality offices containing professional and financial services companies.
- **Ivory Towers:** Characterised by a concentration of Life Sciences and Higher Education Institutions, which employ a highly qualified and professional workforce.

Using the components, it was possible to link back to the original variables through the PCA component loadings (correlations), enabling the identification of the most and least representative town centre characteristics for a given geo-business class. This provides a composite overview of the mix of geo-business characteristics for a given centre. A spider diagram is used to illustrate this mix and provides a comparison of the scores for each geo-business class for 5 London centres in Figure 2.

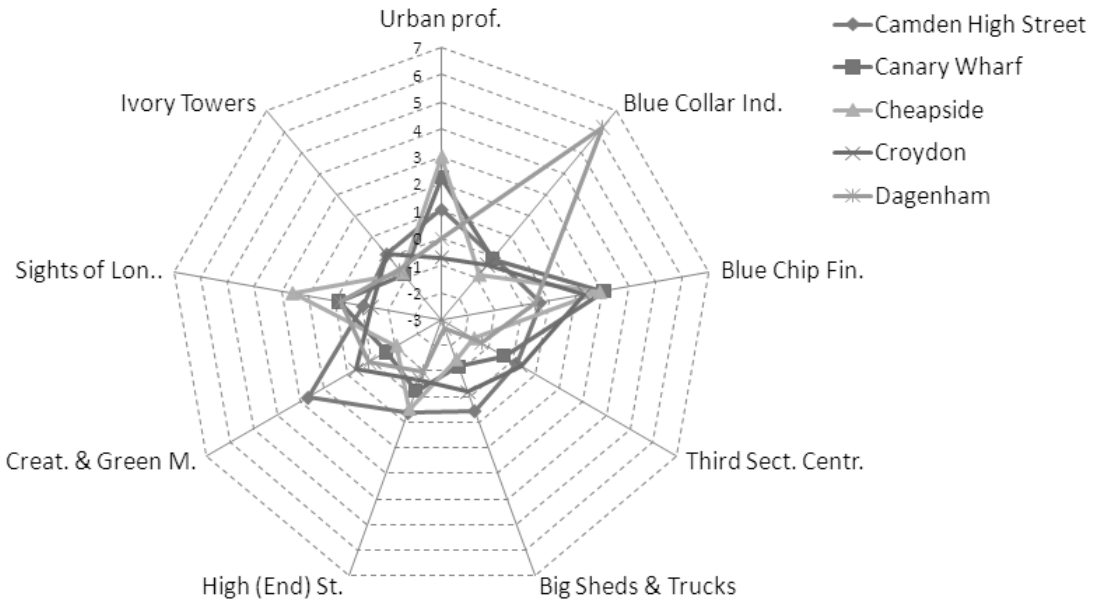



**Figure 2:** Geo-business class scores for a selection of London town centres

## 3. Multi-Criteria Decision Making (MCDM)

Spatial Decision Support Systems rely heavily on Multi-Criteria Analysis methodologies in situations where there is a need for trade-offs, with no apparent optimal solution (Reyck et al. 2005). The MCDM process encompasses a structured approach to develop weights associated to different objectives or criteria, in terms of subjective importance to decision makers, with the overall score of one alternative outcome being the overall performance of the different criteria. Spatial MCDM integrates the multi-criteria decision making methodologies with the spatial database to achieve the "*aggregation and analysis of spatial data and the decision makers preferences into a set of discrete decision alternatives*" (Ghosh 2008).

One potential framework for the resolution of such issues is the Analytical Hierarchy Process (AHP) (Saaty 1990). AHP facilitates decision making by dividing the decision process into a series of pair-wise comparisons, transforming users perceptions and judgements into absolute weights in a hierarchical multilevel decision making framework. These weights are then used to compute and rank decision alternatives. AHP replicates how an individual naturally resolves a multicriteria decision problem, and as such is both a descriptive and prescriptive model of decision making (Forman & Gass 2001).

## 4. Prototype implementation

The potential of the combination of a relevant data base (geo-business classification) and model base (AHP) for business location decision making was explored through a prototype implementation. For the prototype, we implemented a two level decision hierarchy including both the geo-business classification along with some example measures of public transport accessibility. Users were asked to express their location characteristics preferences through pair-wise comparisons in this hierarchy. From these pair-wise comparisons, following the AHP methodology, absolute weights were generated for all decision variables and used to compute the rank of each London neighbourhood. The results were then presented in a mapping interface, allowing the exploration and comparison of

recommended London neighbourhoods, supported by a visualisation of the relevant location variables for each neighbourhood (see Figure 3).

The prototype system was implemented as a light weight web based client, making use of external web services for the mapping interface and diagrams, and a server processing the pair-wise comparisons using the AHP methodology to return to the client the absolute weights and ranking of locations.

To determine the suitability of the chosen methodology and data for solving complex FDI decisions, an evaluation experiment was setup. The participants were 5 expert (FDI promotion professionals) and 5 non expert (UCL academics) users. Three scenarios were designed, based upon real FDI problems Think London have tackled in the past. Each scenario listed information on the company plans for investment into London, including their required location characteristics. Users were asked to "play" the scenarios and use the prototype to find the best location using the information contained in each scenario as guidance.

The primary aim of the user testing process was to evaluate the robustness of the decision making process, i.e. if there was consensus between users on the importance of different location variables. For this, we plotted the average weight or importance users placed on individual geo-business classes per scenario, versus the standard deviation (the variation between users) or disagreement for each variable (see Figure 4) to create a summarised decision matrix.

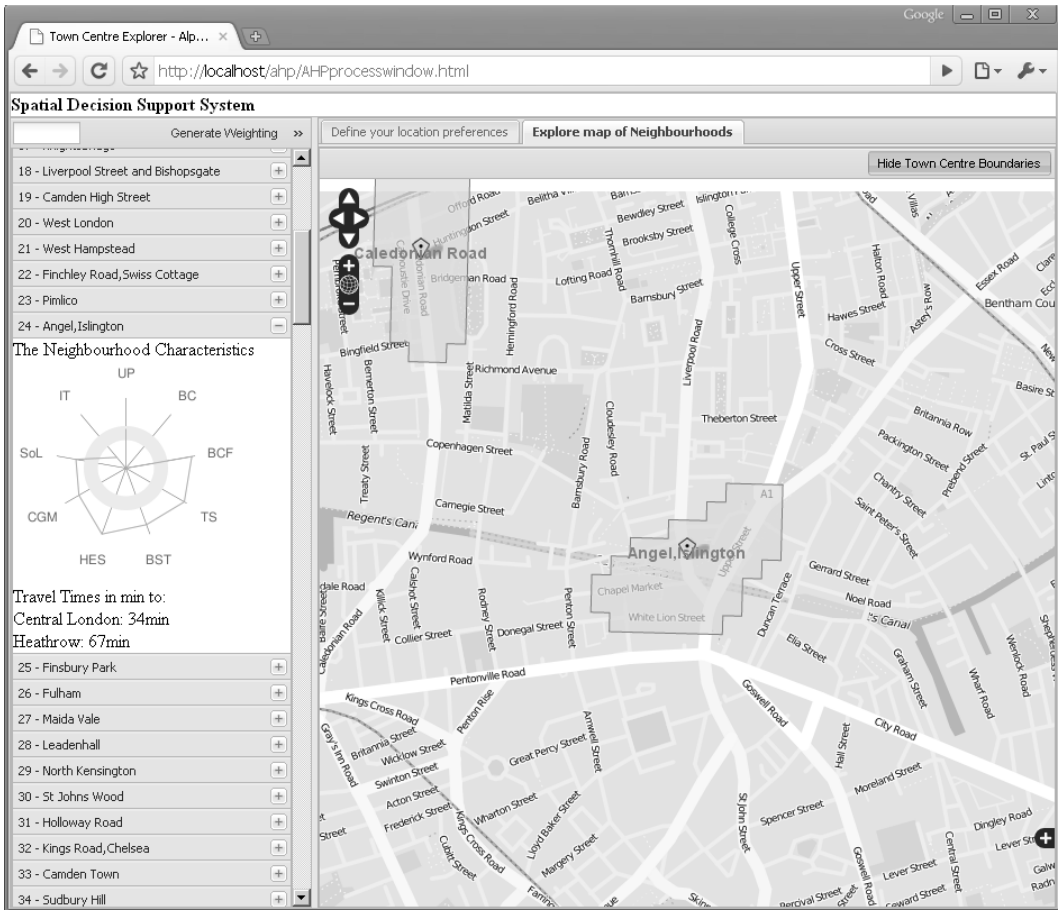



**Figure 3:** Example screenshot of prototype system map interface presenting the results of the process. On the left the user can see the ranked list of recommended locations, with more detailed information about Angel Town Centre.

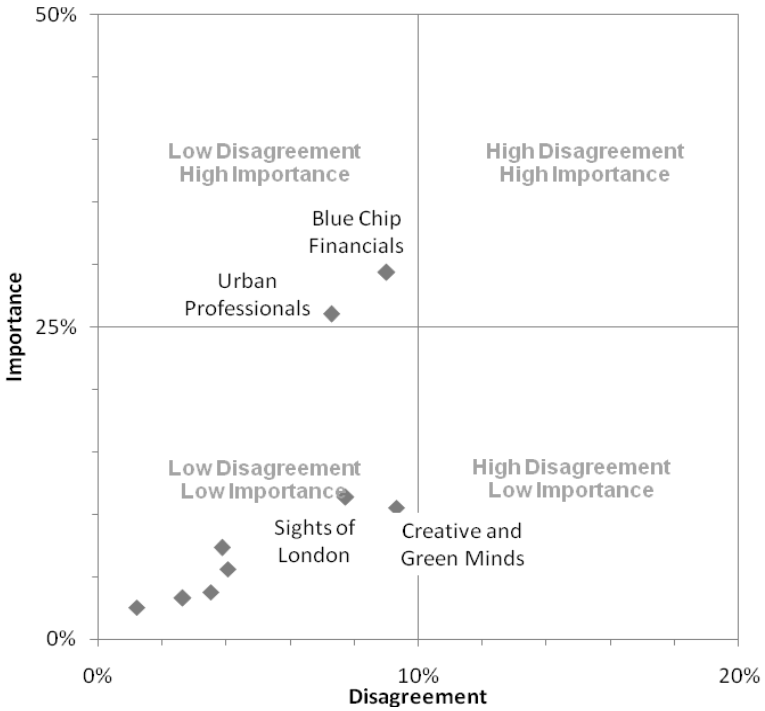



**Figure 4:** Importance/Disagreement matrix for scenario 2. The standard deviation (disagreement) is represented on the X axis, and average score (importance) of decision variables on the Y axis. This example shows relative agreement between users that a mix of "Urban Professionals" & "Blue Chip Financials" type urban environment is desirable.

Preliminary results of the evaluation indicate that the users came to similar conclusions about the weights they attached to neighbourhood characteristics (i.e. how important certain variables were to the scenario). Firstly, there is very clear agreement on the location variables not relevant at all to the decision making. Secondly, we observed that certain sets of neighbourhood characteristics were clearly defined as more important by all users, but the actual weights given resulted in higher disagreement scores. Individual users' judgement on the exact degree of importance of a geo-business class tends to be more contentious than the decision if a geo-business class is relevant or not to the case study. Overall though, the results showed that the decision support methodology was considered robust and consistent producing similar outcomes for each user.

## 5. Conclusion

Through the collection and processing of a set of variables relevant to business location making, a geo-business classification characterising London's diverse neighbourhoods emerged. These profiles, built for each London town centre, can on their own be used to inform and improve business location marketing through better location intelligence, identifying strengths and weaknesses of different areas.

In combination with a relevant MCDM such as AHP, we were able to develop a prototype implementation of a decision environment where users assign their own weights to different area characteristics. The system allows the computation of a suitability index, enabling potential investors to rank centres according to individual needs and presenting a rich environment to iteratively explore, compare and rank London's business centres. The work presented here also contributes to a better understanding of the issues surrounding the integration of previously disparate tools, datasets and methods for spatial decision making. Specifically, we are now confident that AHP makes a significant contribution towards modelling cognitive spatial decision preferences and trade-offs, enabling the delivery of an efficient, relevant and flexible SDSS.

Issues remain though in the procurement and efficient integration of further relevant business location variables not explored in detail so far, such as the integration of live property market information, more detailed industry cluster models, as well as individualised accessibility measures.

## 6. Biography

**Patrick Weber**

*Patrick is currently finishing his Engineering Doctorate thesis looking at decision support tools for FDI promotion activities. He is sponsored by Think London, London's inward investment promotion agency in his work.*

**Dr Dave Chapman**

*Dave is currently Deputy Head of Department of Management Science and Innovation, as well as Course Director for UCL's MSc in Technology Entrepreneurship. Dave has worked as Business Development Manager in UCL's Technology Transfer Office, a Senior Lecturer in UCL's Department of Geomatic Engineering and was the founder and coordinator of the University of London MSc programme in Geographic Information Science.*

## 7. Bibliography

Ackroyd, P., 2001. *London: The Biography* New edition., Vintage.

Carver, S.J., 1991. Integrating multi-criteria evaluation with geographical information systems. *International Journal of Geographical Information Science*, 5(3), 321-339.

Forman, E.H. & Gass, S.I., 2001. The Analytic Hierarchy Process--An Exposition. *OPERATIONS RESEARCH*, 49(4), 469-486.

Ghosh, D., 2008. A Loose Coupling Technique for Integrating GIS and Multi-Criteria Decision Making. *Transactions in GIS*, 12(3), 365-375.

Hebbert, M., 1998. *London: More by Fortune than Design*, London: John Wiley.

Jankowski, P., 1995. Integrating geographical information systems and multiple criteria decision-making methods. *International Journal of Geographical Information Science*, 9(3), 251-273.

Lloyd, D., 2004. *Uncertainty in Town Centre Definition*. PhD Thesis. University College London.

Reyck, B.D., Degraeve, Z. & Grushka-Cockayne, Y., 2005. *System Performance Trade-Offs*, London Business School.

Saaty, T., 1990. How to make a Decision: The Analytical Hierarchy Process. *European Journal of Operational Research*.

Thurstain, M. et al., 2001. *Producing Boundaries and Statistics for Town Centres*, CASA, Center for Advanced Spatial Analysis, London.

Thurstain-Goodwin, M. & Unwin, D., 2000. Defining and Delineating the Central Areas of Towns for Statistical Monitoring Using Continuous Surface Representations. *Transactions in GIS*, 4(4), 305-317.

URBED, 2002. *A City of Villages: Promotion a sustainable future for London's suburbs*, London: Greater London Authority. Available at: http://www.london.gov.uk/mayor/planning/docs/tr11_villages.pdf.

Weber, P. & Chapman, D., 2009. Investing in geography: A GIS to support inward investment. *Computers, Environment and Urban Systems*, 33(1), 1-14.

# Modeling the evolution of metropolitan London using historical GIS

Kiril Stanilov

Centre for Advanced Spatial Analysis (CASA), University College London, 1-19 Torrington Place
London WC1E 7HB UK
Tel: +44(0)20 7679 1812 | Email: kiril.stanilov@ucl.ac.uk

KEYWORDS: historical GIS, urban modeling, metropolitan form, land use change

## 1. Introduction

The paper presents the results of a study that traces the emergence and evolution of London's metropolitan form from the second half of the 19th century to the present. A longitudinal land use database, developed by digitizing historical Ordnance Survey maps, is used to build a constrained cellular automata model, which explores key factors shaping the patterns of London's urban growth. The underlying goal of this research is to address a chronic weakness of urban modeling concerning deficiencies in the quality of data used to build urban growth models (Longley and Mesev, 2000). This goal is pursued by the development of a unique database that stretches the boundaries of conventionally employed data in urban modeling along several dimensions.

## 2. Data and methodology

The development of the database used in this study is driven by several key principles. First, perhaps less unique but very important for the analysis of metropolitan form generation, is the large spatial extent of the study area which covers 200 sq km stretching from Hyde Park westward to the edges of London's Green Belt (Figure 1). This area was selected to explore a major axis of London's historic growth, covering one of the most dynamically evolving segments of London's metropolitan fabric.

Second, the study uses highly detailed historic Ordnance Survey (OS) maps at a scale of 1:2,500, which allows the identification of a wide range of land use categories and building types with a high level of spatial and interpretational accuracy. This presents a significant improvement in data resolution compared to traditional land use and land cover databases derived from remotely sensed (RS) imagery. Most historical land use change studies rely on RS data with a pixel resolution in the 10-to-30 m range (limited by the resolution of historical LANDSAT and SPOT satellite images), which presents significant challenges in the identification of various urban land use classes. In most such cases a distinction is made between residential, non-residential, and recreational uses, but even these broad categorizations are vulnerable to inaccuracies of interpretation due to the low resolution of the satellite images. In contrast, the fine-scale OS maps allow the precise identification of close to 60 land use classes and building types with an accurate representation of actual parcel boundaries. In this study a manual interpretation of the maps served as a basis for the creation of GIS polygons of homogeneous land uses delineated by parcel boundaries. Residential, mixed use, retail, and office classes were assigned through a process involving the interpretation of building footprints; verification of building type in Microsoft Virtual Earth and Google Street View; and cross-referencing the results with several land use databases for Greater London. As the majority of the non-residential buildings are clearly labeled in the OS maps (see Table 1, column 3), these OS designations were directly used for the assignment of land use in aggregated classes (Table 1, column 2). Thus land use polygon coverages were created for all of the seven study periods described in the next paragraph. The vector-based maps were then converted to grid coverages with a cell size of 25 by 25 meters. These fine-grained lattices are then used in the development of a cellular automata urban growth model based on a more precise spatial accuracy and a refined identification of land use interactions. Such method of database development eliminates the problems associated with cell heterogeneity encountered in models based on lower resolution data sources. The project experiments

with different levels of land use aggregation to determine an optimal number of classes generating most accurate model outcomes.

A third unique feature of the database employed by this project is the extensive time coverage, which spans the last 130 years of London's urban growth. The developed data set includes time series land use maps showing the evolution of metropolitan form in 20–year increments, starting from 1875 onwards (Figure 2). The time coverage of the project stands in marked contrast with conventional urban modeling techniques, which rarely go back in time more than a decade or two. The unusually long time horizon of this study (reaching the pre-urban era of the metropolitan periphery) allows to trace the emergence and evolution of London's suburban fabric from its incipient stages of urbanization.

## 3. Analysis

The high resolution of the data, the extensive longitudinal coverage, and the large areal extent of the study area serve an important purpose. They allow the development of a model that reveals the emergence of London's metropolitan spatial structure as a result of myriad capillary changes that have taken place on the level of individual parcels over the course of nearly one and a half centuries. The application of a cellular automata model, used in this study, is viewed as particularly suitable to capture such fine-grain local transformations of the built environment, demonstrating how local action generates global patterns (Batty, 1997). Unlike conventional aggregative modeling techniques, CA-based models can handle high-resolution applications easily in conjunction with the precision of high-quality data sets typically resident in GIS (White and Engelen, 2000).

A key characteristic of the CA model developed in this study is that it is used not for prediction of future land use but for analysis of the determinant factors of metropolitan growth patterns in a historical perspective. The digitized land use maps are used to calculate growth rates by land use class, which are then fed into the model as constraints. The process of calibration of model parameters is used to explore the degree of influence on urban growth patterns of the following key factors: 1) distance to the metropolitan center; 2) distance to the major elements of the transportation network (roads, waterways, railways, tramways, underground and railway stations); 3) distance to existing suburban centers/nuclei; and 4) the forces of attraction and repulsion between the various land use categories. The temporal scale of the study allows tracing the dynamic variations in the impact of these factors through time (Figure 3).

The process of model calibration is aided by analysis of the spatial behavior of the identified land use classes performed in ArcGIS (e.g., measuring the distribution of various residential building types in relation to the location of railway stations, distance to metropolitan center, proximity to other land use classes, etc.). The results are used to inform parameter calibration and to derive land use change potential for all of the cells in the model for all of the seven time periods (1875, 1895, 1915, 1935, 1960, 1985, and 2005). The initial year for running the model is 1875 with data based on the historical land use map from that year. The model is updated in increments of one year and the results are compared to the historical changes documented from the OS maps. Thus adjustments in the model calibration are made in points in time when recalibration of the initial parameters is required.

The model utilizes the advantages of CA approaches for exploration of complex spatial phenomena such as the dynamic formation of patterns of nucleation, diffusion, and conurbation within the metropolitan fabric (Couclelis, 2002). The study illuminates the importance of the pre-urban development patterns (agricultural land ownership, network of country roads, existing settlement structure) as major spatial determinants framing the subsequent patterns of urban growth (Figure 4 and 5).

## 4. Conclusions

The use of high-resolution cadastral data holds a promise for significant qualitative improvements in the fidelity of representing urbanization patterns and processes in the modeling world. While the availability of historical cadastral data is still a rarity in most institutional contexts, this situation is changing rapidly with the progress of the digital revolution. Urban modelers are charged to apply forward thinking, which should not be confined by the limitations of the data and concepts in currency today. Improving the level of realism in representing urban environments can lead not only to an enhanced comprehension of model design and outcomes, but to an enhanced theoretical and empirical grounding of the entire field of urban modeling.

## 5. Tables and figures

**Table 1.** Land use classification based on historical OS maps

| Land Use Categories | Recorded Land Use Classes | OS designations |
|---|---|---|
| residential | mid and low-rise apartments | |
| | high-rise apartments | |
| | Cottage houses | |
| | Terrace houses | |
| | Mews | |
| | semi-detached houses | |
| | small lot detached houses | |
| | detached houses | |
| | Lodges | lodge, inn, hotel |
| commercial | old fabric | |
| | general | |
| | Retail | store |
| | entertainment | public houses, cinemas, dance halls |
| | Garages | garage |
| office | Office | |
| institutional | institutional public | city hall, library, police station, post office, fire station, hospital, military, prison |
| | educational | daycare, school, college, university |
| | religious | church, convent, rectory, priory, friar, hermitage, vestry, synagogue, temple |
| | Stadia | stadium |
| industrial | industrial | works, mill, wharf, dock, depot, brewery, malthouse, laundry |
| | Utilities | sewer works, gas works, water works, reservoir |
| transportation | Roads | |
| | Railway lines | |
| | Railway stations | station |
| | Rivers | |
| | Canals | |
| | airports | airport |
| open space | Estates | house |
| | Parks | park, green |
| | recreation | golf course, athletic club, recreation ground, playground, shooting range |
| | cemeteries | cemetery, burial ground |
| | allotment gardens | allotments |
| | nurseries | nursery |
| farms | Farms | farm |


**Figure 1.** Study area

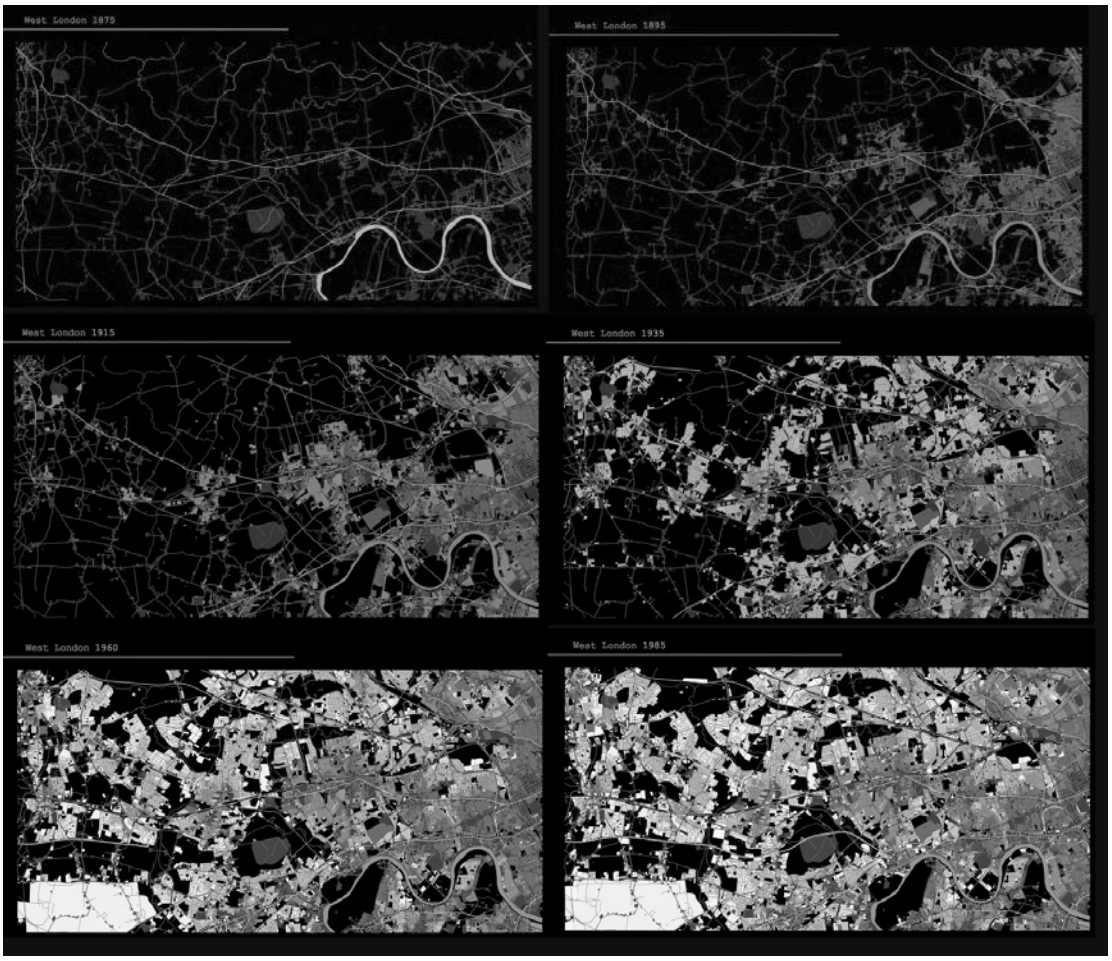



**Figure 2.** Land use growth patterns in West London, 1875-2005

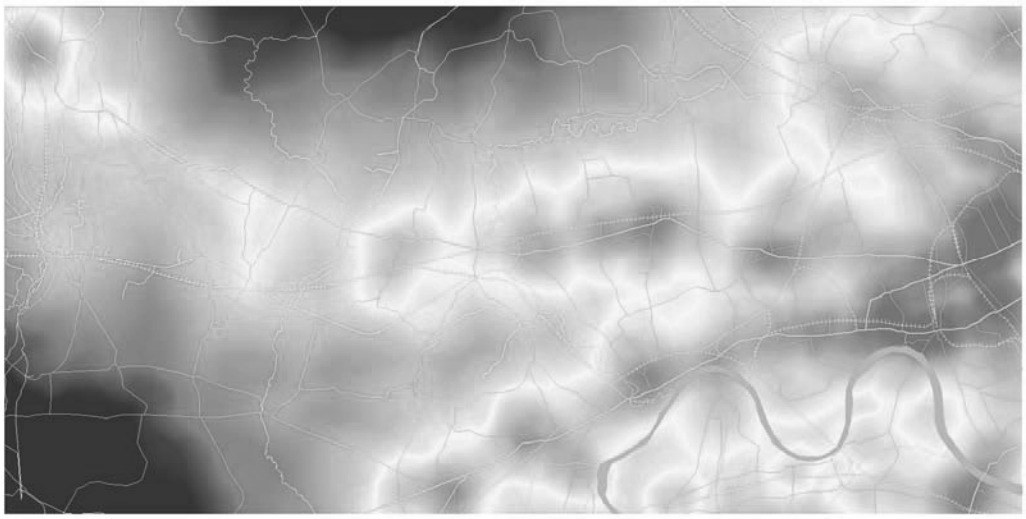


**Figure 3.** Composite accessibility in 1875 based on distance to major roads, railway stations, city centre, and suburban nuclei.

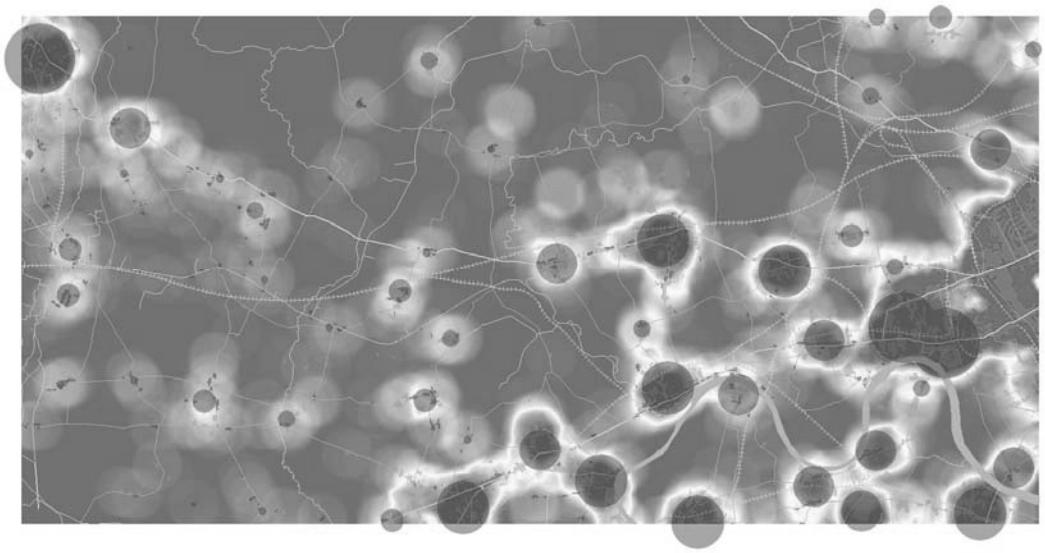


**Figure 4.** Identification of historical suburban clusters in 1875.

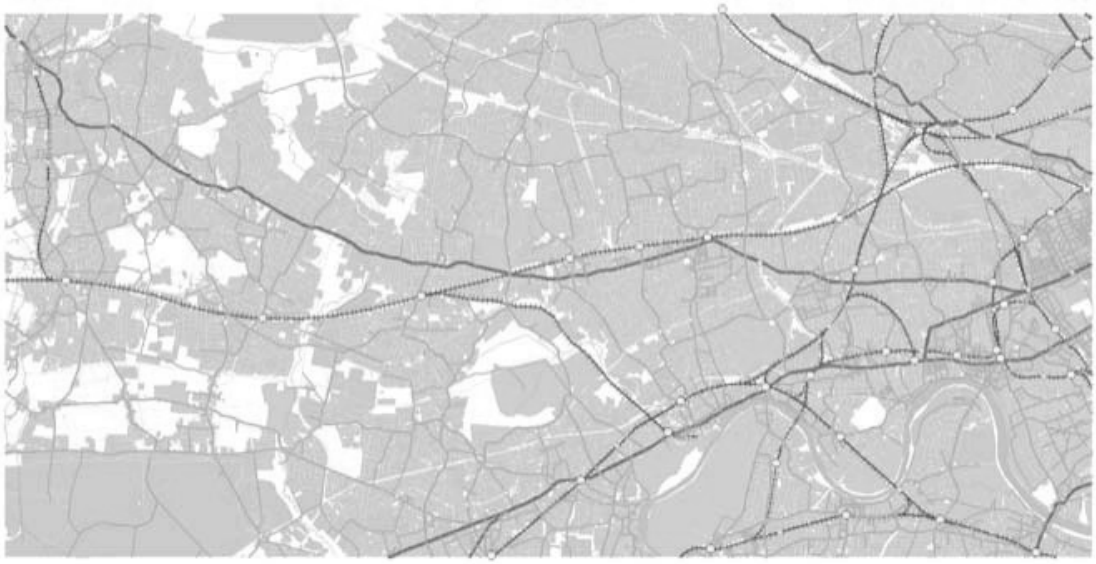


**Figure 5.** Infrastructure (roads, canals, railways and stations) from 1875 still existing in 2005.

## 6. References

Batty M (1997) Cellular automata and urban form: a primer, *Journal of the American Planning Association* 63(2) pp266-274.

Couclelis H (2002) Modeling frameworks, paradigms, and approaches, in: Clarke, K.C., Parks, B.E., Crane, M.P. (eds.), *Geographic Information Systems and Environmental Modeling*, Longman & Co, New York, pp. 36–50.

Longley PA and Mesev V (2000) On the measurement and generalisation of urban form, *Environment and Planning A* 32 pp473-488.

White R and Engelen G (2000) High-resolution integrated modelling of the spatial dynamics of urban and regional systems, *Computers, Environment and Urban Systems* 24 pp383-400.

## 7. Acknowledgements

The project is funded by the European Commission's Framework Programme 7 through a Marie Curie fellowship.

## Biography

The author is a Marie Curie Research Fellow at the Centre for Advanced Spatial Analysis at UCL. Kiril Stanilov's research interests are centered on explorations of contemporary patterns of urban growth and change, and the role played by public policies in shaping urban form transformations.

# A geodemographic classification of London primary schools

Anne Gibbs, John Stillwell and Linda See

School of Geography, University of Leeds, Leeds, West Yorkshire, LS2 9JT, UK
Tel: +44 (0)113 34 3300
Email: geo5a2eg@leeds.ac.uk ; j.stillwell@leeds.ac.uk ; l.m.see@leeds.ac.uk

KEYWORDS: primary schools, geodemographic classification, SATs results, cluster analysis.

## 1. Introduction

Since the introduction of the Pupil Level Annual School Census (PLASC) in 2002, much research has considered the nature of the relationships between schools, their neighbourhoods and the performance of their pupils (see, for example, Butler et al., 2007; Goldstein et al., 2007; Gordon & Monastiriotis, 2007; Hamnett et al., 2007; Webber & Butler, 2007) although only a limited amount of it has focussed on the primary sector (e.g. Goldstein et al., 2006; Harris & Johnston, 2008; Strand & Demie, 2006). The density and diversity of the pupil population in London creates a particularly complex pattern of relationships between primary schools and their neighbourhoods. One way of summarising these relationships is by classifying schools according to the key ethnic and socio-economic characteristics of their pupil populations. This not only enhances understanding by highlighting the similarities and differences between the schools, but such a classification also has the potential to be used by managers and policy makers as a benchmarking tool.

The purpose of this paper is to present a classification of state maintained London primary schools. The methodology is explained in section 2, followed by the resultant classification in section 3. Section 4 considers how the classification might be used to explain differentials in school performance.

## 2. Methodology

A database comprising the key socio-economic and ethnic characteristics of the pupil populations of London primary schools was constructed from the School Census of May 2007, held by the Department for Children, Schools and Families (DCSF) within the National Pupil Database. The following nine input variables were selected to derive the classification of schools:

Ethnicity variables
% African
% Caribbean
% Pakistani
% White British
% Any Other White

Socio-economic variables
% eligible for Free School Meals (FSM)
% with English as an Additional Language (EAL)
% with Special Educational Needs (SEN) (all levels of need including Statemented pupils)
% mobile (i.e. % of 2006/7 Year 6 pupils who entered their school after the beginning of Year 5)

Note: For all variables except % mobile, the base population is the total number of pupils in the statutory years of Primary education: Reception to Year 6.

The particular ethnicities selected comprise the largest ethnic groupings within the pupil population and, more significantly, those that have been found to be the most residentially and educationally segregated (Johnston et al., 2006). The socio-economic variables were selected on the basis that they reflect fundamental characteristics of the school population which research has shown are likely to have an impact on schools' Key Stage 2 SATs results (section 5). They are in line with both the National Foundation for Education Research (NFER) (Benton et al., 2007) and Contextual Value Added (CVA) models (DCSF, 2007a), as well as the DCSF's own 'Primary Families of Schools' classification (DCSF, 2008). The variables were standardised using range standardisation, but no weighting was applied.

Using the selected variables, the classification was derived using the k-means algorithm in SPSS. This iterative method was favoured primarily because the schools database is not inherently hierarchical and therefore not naturally suited to a hierarchical clustering method. Furthermore, it was considered that an optimal solution is more likely to be reached, providing sufficient iterations are conducted. The k-means algorithm, one of the most commonly used methods in geodemographic classification, was used in the Census area classifications, in conjunction with Ward's clustering method (Office for National Statistics, 2004), and also by Experian for the MOSAIC classification (Harris et al., 2005). The principal drawback of this method is that the number of clusters has to be pre-determined. However, this was overcome by running the k-means clustering algorithm for k=4 through to k=22 clusters and then assessing the relative merits of the resultant classifications. The ultimate aim of cluster analysis is to define clusters which are distinct and where the individual cases are tightly concentrated within each cluster. Thus the alternative solutions were assessed in a two-stage process. Firstly, all the solutions were compared using the following two measures:

1. average distance from cluster centre (using the standardised data) to indicate compactness of the clusters, and
2. average difference from mean cluster membership to indicate the evenness of cluster size.

Secondly, the four most favourable solutions from the first stage of the assessment were further analysed for the homogeneity and separateness of their clusters, using the raw data and the three measures of cluster validity detailed below, viz. Silhouette Index; Davies-Bouldin Index and average distance to cluster centre. The results of the first stage of the assessment are displayed in Figure 1. Step reductions in both measures suggest that the k=7, k=10, k=14 and k=16 cluster solutions are the most favourable.

The results of the second-stage assessment are shown in Table 1. Three measures of cluster validity have been used:

- the Silhouette Index (SI) (Rousseeuw, 1987)[1], which measures how appropriately each individual case is assigned to its cluster: once the 'silhouette width' of each individual case has been calculated, these can then be aggregated and averaged at the cluster and dataset levels to reflect the compactness and separateness of the clusters (the higher the SI the better the clustering solution);

[1] Suppose *a* represents the average distance of a case from all other cases in its cluster and *b* represents the minimum of the average distances of the case from the cases of the other clusters, the silhouette width *s* of the case is defined as:

$$s = \frac{b - a}{\max [a,b]}$$

The value of *s* varies from -1 to 1. The closer the value is to 1, the better clustered the case is. Conversely, a value close to -1 means that the case is 'misclassified' and actually lies somewhere in between the clusters. The SI is the average silhouette width of all the cases in a cluster or across all clusters..

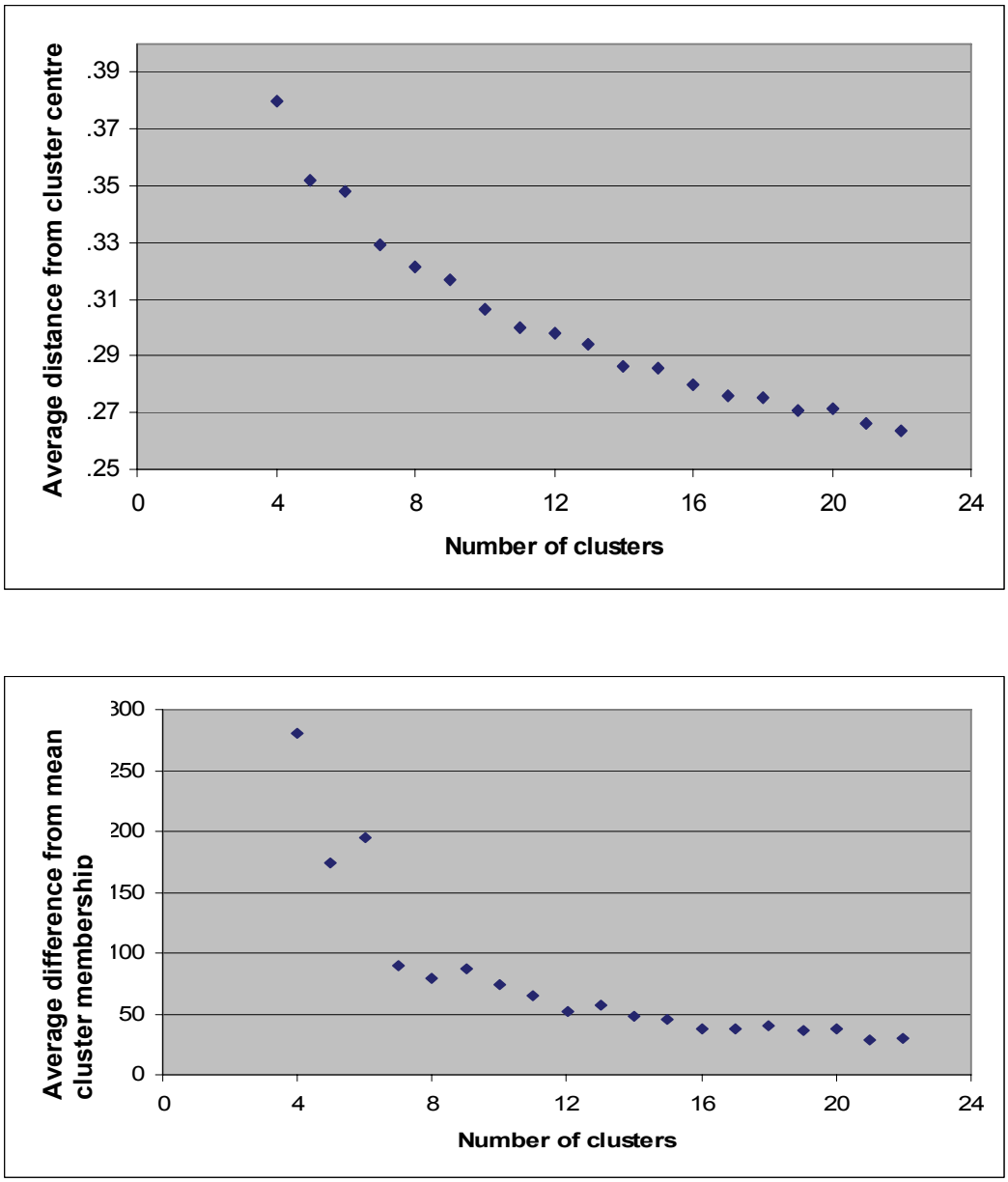



**Figure 1.** Evaluating alternative cluster solutions

- the Davies-Bouldin Index (DBI) (Davies and Bouldin, 1979)[2] is a function of the ratio of within-cluster distances to between-cluster separation: clusters with desirable characteristics will have low values;

[2] The DBI is calculated as follows: For each pair of clusters: if $a_i$ is the average distance of all cases in cluster $i$ to the centre of that cluster, and $b_{ij}$ is the distance between clusters $i$ and $j$, calculate:

$$R = \max_{i \neq j} \frac{a_i + a_j}{b_{ij}}$$

$$\text{Then, DBI} = \frac{1}{k} \sum^{n} R$$

- Average Distance to Cluster Centre is a more longstanding measure of cluster validity and is an entirely internal measure which does not take account of distances between clusters. It is therefore limited in terms of providing an overall assessment of the cluster solution but, nevertheless, still provides a useful measure of the compactness of the clusters.

**Table 1** Evaluation of favoured clustering solutions

| Number of clusters | Silhouette Index (high preferable) | Davies-Bouldin Index (low preferable) | Average Distance to Cluster Centre (low preferable) |
|---|---|---|---|
| 7 | 0.362 | 2.461 | 51.975 |
| 10 | 0.316 | 4.955 | 77.298 |
| 14 | 0.307 | 1.582 | 22.499 |
| 16 | 0.260 | 4.506 | 22.115 |

Considering the results of this second-stage assessment, the 14-cluster solution has the most favourable set of characteristics overall and therefore was selected as the final classification. The clusters in this particular analysis converged after 32 iterations.

### 3. The Classification of London Primary Schools

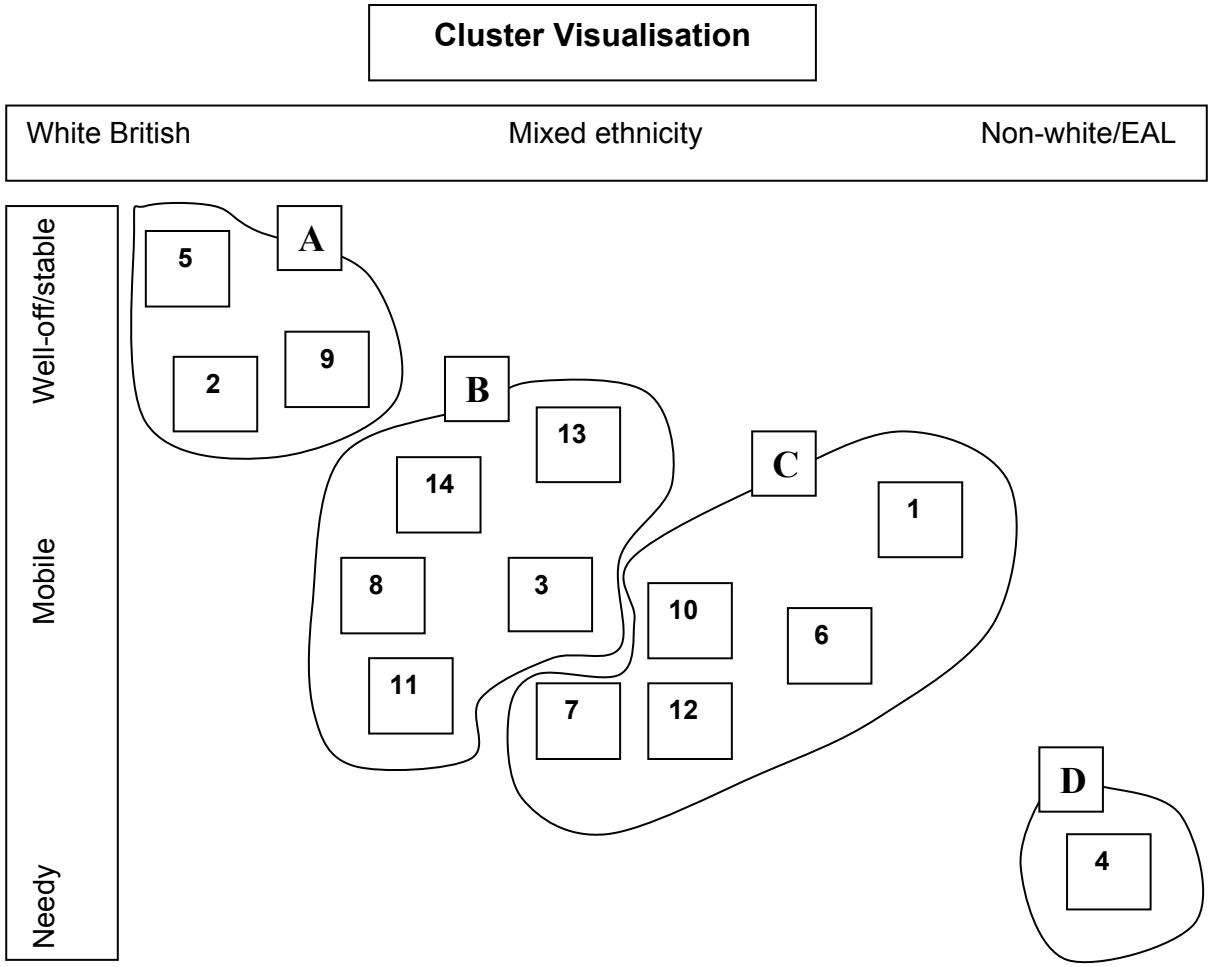
The chosen clustering solution is graphically represented in Figure 2. In this diagram, the clusters are arranged to reflect, as closely as possible, the relative distances between them in multi-variate space, whilst also indicating their placement on the horizontal 'ethnicity' axis and the vertical 'socio-economic' axis.

Clusters 4 and 5 can be regarded as the 'polar opposites' of the classification. Cluster 4 comprises a very distinct group of 53 schools which have predominantly very needy Bangladeshi pupil populations. The vast majority (89%) of these schools are in the borough of Tower Hamlets. Cluster 5, on the other hand, is the largest cluster comprising 246 schools (15.5% of the total) characterised by non-mobile White British pupils from well-off families. The great majority of these schools are located in the suburbs, particularly in the south and east of the capital. The other 12 clusters lie between these two extremes and contain schools with differing ethnic compositions and varying degrees of neediness and mobility within their pupil populations (Table 2).

Overall, the classification explains 66.8% of the variance in the dataset, but rather more of the variance in ethnicity (71.5%) than in socio-economic factors (58.4%). The variables least explained by the classification are mobility (36.3%) and SEN (37.6%). This is hardly surprising, as both these factors are likely to be related to a range of external variables such as, for mobility, household moves and employment changes as well as international migration flows and, for SEN, genetic characteristics, parenting, early childhood trauma, etc.

As it can be difficult to accurately characterise as many as 14 clusters, the following four super-groups have been discerned:

A. Schools with predominantly well-off White British pupils (Clusters 5, 2 and 9);
B. Schools with ethnically-mixed European and second generation African and Caribbean migrant pupil populations (Clusters 14, 8, 3, 13 and 11);
C. Schools predominantly controlled by the local education authority (LEA) with high proportions of first generation South Asian migrant pupils (Clusters 1, 6,7, 10 & 12);
D. Schools with predominantly needy Bangladeshi pupil populations (Cluster 4).




Source: After Harris et al. (2005) Figure 6.3, p.170.

**Figure 2.** Cluster visualisation

Although Groups B and C are intertwined in the visualisation in Figure 2, in reality there are distinct differences between them (Table 2). Whilst Group C schools are predominantly LEA-controlled and have relatively high proportions of children with EAL (English as an Additional Language) and/or who are eligible for free school meals (FSM), Group B schools are much more likely to be externally controlled VA (Voluntary Aided) or foundation schools and have higher proportions of relatively stable Afro-Caribbean and White children, albeit relatively needy ones.

There is a clear pattern in the spatial distribution of most of the clusters (Figure 3), largely determined by the predominant ethnicity (the City of London has been excluded as there is only one primary school within its boundaries). Group A clusters are particularly concentrated in the outer London Boroughs, except to the north. Group B clusters are concentrated particularly in the inner London Boroughs to the south of the Thames, whilst Group C clusters are found throughout inner London and the northern suburbs, but especially in Newham and Brent. Not being associated with particular ethnicities, the schools in the needy multi-ethnic Cluster C7, with its relatively high percentage of SEN pupils, and Cluster C10, with its high mobility rates, are dispersed throughout this area. The predominant characteristics of the clusters are summarised in Table 3.

**Table 2.** Cluster characteristics (indexed to global average = 100)

| Cluster | No. of schools | LEA-controlled schools | VA / foundation schools | AFRICAN | CARIBBEAN | INDIAN | BANGLADESHI | PAKISTANI | WHITE BRITISH | ANY OTHER WHITE | FSM | EAL | SEN | MOBILE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A2** | 148 | 129 | 35 | 56 | 41 | 44 | 19 | 39 | 177 | 56 | 102 | 40 | 126 | 111 |
| **A5** | 246 | 85 | 133 | 25 | 17 | 44 | 6 | 19 | 209 | 47 | 25 | 18 | 66 | 64 |
| **A9** | 173 | 71 | 165 | 35 | 38 | 70 | 15 | 45 | 145 | 137 | 36 | 54 | 69 | 72 |
| **B3** | 73 | 38 | 240 | 343 | 158 | 21 | 25 | 10 | 38 | 61 | 118 | 139 | 104 | 77 |
| **B8** | 111 | 82 | 140 | 144 | 374 | 47 | 23 | 74 | 47 | 76 | 120 | 88 | 115 | 97 |
| **B11** | 140 | 115 | 67 | 132 | 93 | 49 | 96 | 61 | 94 | 96 | 160 | 106 | 136 | 97 |
| **B13** | 60 | 60 | 189 | 64 | 78 | 60 | 37 | 29 | 67 | 310 | 75 | 123 | 75 | 88 |
| **B14** | 184 | 89 | 125 | 95 | 128 | 102 | 27 | 116 | 91 | 96 | 77 | 88 | 93 | 95 |
| **C1** | 76 | 139 | 13 | 75 | 57 | 793 | 85 | 423 | 22 | 43 | 82 | 194 | 91 | 111 |
| **C6** | 34 | 136 | 19 | 88 | 81 | 216 | 160 | 929 | 25 | 90 | 118 | 190 | 117 | 113 |
| **C7** | 126 | 130 | 33 | 193 | 197 | 49 | 87 | 52 | 30 | 201 | 181 | 159 | 140 | 126 |
| **C10** | 71 | 126 | 41 | 131 | 97 | 130 | 52 | 132 | 51 | 133 | 133 | 155 | 114 | 260 |
| **C12** | 97 | 121 | 54 | 142 | 84 | 114 | 258 | 165 | 37 | 109 | 161 | 180 | 105 | 112 |
| **D4** | 53 | 131 | 31 | 43 | 16 | 14 | 1425 | 32 | 19 | 29 | 204 | 216 | 92 | 91 |
| **Total/ Global Average** | **1592** | **69.2** | **30.8** | **13.0** | **6.9** | **4.3** | **5.2** | **3.1** | **37.5** | **9.3** | **28.1** | **39.2** | **23.2** | **11.6** |

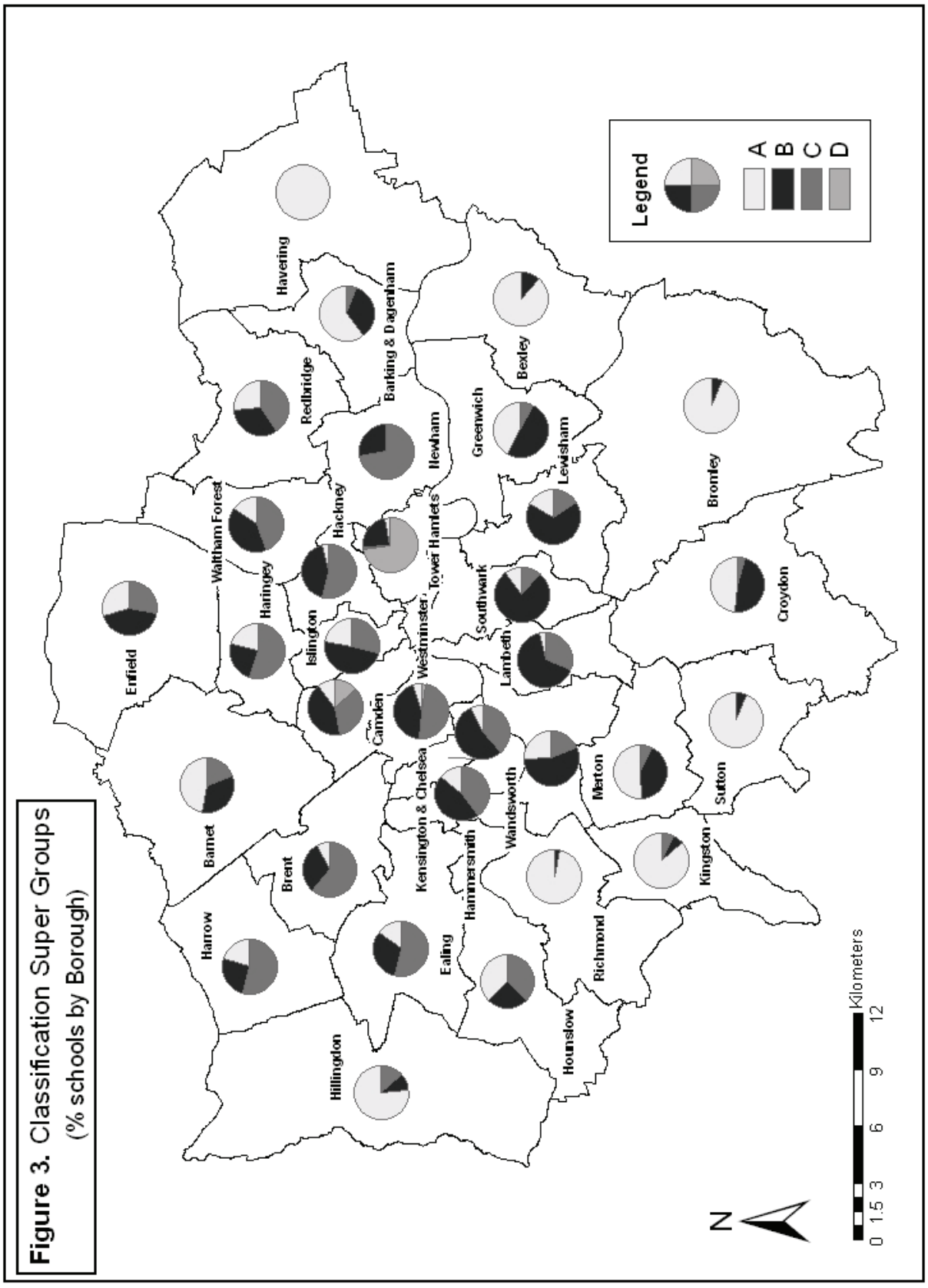



**Figure 3.** Classification Super Groups (% schools by Borough)

**Table 3.** Key cluster characteristics

| Cluster | Predominant school characteristics | No. of Schools |
|---|---|---|
| **A2** | Predominantly White British, socio-economically average but above-average proportion of SEN, predominantly LEA-controlled | 148 |
| **A5** | White British, well-off, above-average number of VA/foundation schools | 246 |
| **A9** | Predominantly White British and Other White, well-off, above-average number of VA/foundation schools | 173 |
| **B3** | Concentration of African pupils, mainly VA/ foundation schools | 73 |
| **B8** | Afro-Caribbean, above-average VA/foundation schools | 111 |
| **B11** | Multi-ethnic, needy, relatively high rate of SEN, majority LEA-controlled | 140 |
| **B13** | Well above average proportion Other White, relatively well-off, majority VA/ foundation, below average FSM | 60 |
| **B14** | Multi-ethnic but relatively well-off with below average FSM & EAL, mixed governance | 184 |
| **C1** | Asian, high rate of EAL, predominantly LEA-controlled | 76 |
| **C6** | Asian, poorer, high proportion of EAL, predominantly LEA-controlled | 34 |
| **C7** | Mixed ethnicity, relatively high mobility & SEN rates, needy, predominantly LEA schools | 126 |
| **C10** | Multi-ethnic, very high mobility rate, predominantly LEA-controlled | 71 |
| **C12** | Multi-ethnic, high EAL, stable but needy, majority LEA-controlled | 97 |
| **D4** | Bangladeshi, very needy, predominantly LEA-controlled | 53 |

## 4. School Performance

Table 4 ranks the clusters according to the average percentage of high performing pupils which they have (i.e. those attaining above the national target level in the KS2 tests), set in the context of the socio-economic characteristics and the deprivation rates of the schools' pupil populations. The highest performing clusters are the better-off White dominated Clusters A9 and A5. Schools in the other White-dominated Cluster (A2) perform much less well, probably reflecting the relatively high proportion of pupils claiming FSM and with SEN in their populations.

The Bangladeshi schools of Cluster D4, on the other hand, seem to perform unexpectedly well considering the particularly high rates of FSM and IDACI (Income Deprivation Affecting Children Index)[1] amongst their pupils. It is difficult to explain this finding, except to speculate that their ethnically exclusive pupil population may in some way enhance performance.

**Table 4.** Contextual performance of schools by cluster (mean percentages)

| Cluster | High performers | Eligibility for FSM | EAL | Mobility rate | SEN | IDACI rate |
|---|---|---|---|---|---|---|
| **A9** | 146 | 36 | 54 | 72 | 69 | 60 |
| **A5** | 137 | 25 | 18 | 64 | 66 | 43 |
| **B13** | 126 | 75 | 123 | 88 | 75 | 93 |
| **B14** | 112 | 77 | 88 | 95 | 93 | 92 |
| **C1** | 105 | 82 | 194 | 111 | 91 | 78 |
| **D4** | 86 | 204 | 216 | 91 | 92 | 179 |
| **B3** | 84 | 118 | 139 | 77 | 104 | 139 |
| **B8** | 80 | 120 | 88 | 97 | 115 | 122 |
| **A2** | 78 | 102 | 40 | 111 | 126 | 99 |
| **C12** | 76 | 161 | 180 | 112 | 105 | 138 |
| **C6** | 76 | 118 | 190 | 113 | 117 | 109 |
| **C10** | 74 | 133 | 155 | 260 | 114 | 106 |
| **C11** | 71 | 160 | 106 | 97 | 136 | 134 |
| **C7** | 63 | 181 | 159 | 126 | 140 | 149 |
| **Total** | **32.2** | **28.1** | **39.2** | **11.6** | **23.2** | **32.9** |

## 5. Conclusion

Although the focus here has been on the differences in performance between the clusters, the research also revealed evidence of ethnic and social sorting within primary schools. This is in line with the findings of Johnston et al. (2006) which indicate more ethnic segregation in primary schools than in neighbourhoods.

To conclude, the classification presented here enhances our understanding of the complex and diverse range of state maintained primary schools in London by highlighting the similarities and differences between them. Inevitably, there is much more analysis which could be done. Geodemographic analysis has its critics (Vaos & Williams, 2001) but, as Harris et al. (2007) have recently suggested, geodemographic classifications remain potentially valuable as a framework for analysis. Thus the classification presented here has potential to provide the context for more detailed analyses of ethnic and social disparities between schools and their relationship to educational attainment, but also as a benchmarking tool by managers and policy makers.

## 6. Acknowledgements

The data used in this analysis is from the PLASC and KS2 Tables of the National Pupil Database held by the DCSF.

## References

Benton, T. Chamberlain, T., Wilson, R. & Teeman, D. (2007) '*The Development of the Children's Services Statistical Nearest Neighbour Benchmarking Model: final report'*, NFER, Slough.

Butler, T., Hamnett, C., Rumsden, M. & Webber, R. (2007) 'The best, the worst and the average: secondary school choice and education performance in East London', *Journal of Education Policy*, 22(1), 7-29.

DCSF, (2007), '*A Technical Guide to Contextual Value Added 2007* Model', downloaded from the DCSF website on 18th June 2008
http://www.dcsf.gov.uk/performancetables/primary_07/2007GuidetoCVA.pdf

DCSF (2008) '*Primary Families of Schools: London Primary Schools'*, June.

Goldstein, H., Burgess, S. & McConnell, B. (2006) 'Modelling the effect of pupil mobility on school differences in educational achievement', *Working Paper 06/156*, Centre for Market and Public Organisation, University of Bristol, September.

Goldstein, H., Burgess, S., & McConnell, B. (2007) 'Modelling the effect of pupil mobility on school differences in educational achievement', *Journal of the Royal Statistical Society A*, 170(4), 941-954.

Gordon, I. & Monastiriotis, V., (2007) 'Education, Location, Education: A Spatial Analysis of English Secondary School Public Examination Results'*, Urban Studies*, 44(7), 1203-1228

Hamnett, C, Ramsden, M. & Butler, T. (2007) 'Social Background, Ethnicity, School Composition and Educational Attainment in East London, *Urban Studies*, 44(7), 1255-1280.

Harris, R., Sleight, P., Webber, R., (2005) '*Geodemographics, GIS and Neighbourhood Targeting'*, Wiley, Chichester.

Harris, R., Johnston, R. & Burgess, S., (2007) 'Neighbourhoods, Ethnicity and School Choice: Developing a Statistical Framework for Geodemographic Analysis', *Population Research Policy Review*, 26, 553-579.

Harris, R. & Johnston, R. (2008) 'Primary Schools, Markets and Choice: Studying Polarization and the Core Catchment Areas of Schools', *Applied Spatial Analysis*, 1, 59-84.

Johnston, R., Burgess, S., Wilson, D. & Harris, R. (2006) 'School and Residential Ethnic Segregation: An Analysis of Variation Across England's Local Education Authorities', *Regional Studies* 40(9), 973-990.

Power, S., Warren, S., Gillborn, D., Clark, A., Thomas, S. & Coate, K. (2002) 'Education in Deprived Areas: Outcomes, Inputs and Processes', *Perspectives on Education Policy*, Institute of Education, University of London, London.

Social Exclusion Unit (2001) '*A New Commitment to Neighbourhood Renewal: National Strategy Action Plan'*.

Strand, S., Demie, F., (2006) 'Pupil mobility, attainment and progress in primary school'*, British Educational Research Journal,* 32(4), 551-568.

Voas, D. & Williamson, P. (2001) 'The diversity of diversity: a critique of geodemographic classification', *Area*, 33(1), 63-76.

Webber, R. & Butler, T. (2005) 'Classifying pupils by where they live: how well does this predict variations in their GCSE results?' *CASA (Centre for Advanced Spatial Analysis) Working Paper 99*, University College London, December.

## Biographies

*Dr Anne Gibbs has recently completed an MSc in GIS at the University of Leeds. She now plans to further her research in the geography of education, particularly in relation to the ethnic and socio-economic differentials between pupil populations. She works part-time as a teaching assistant in an inner-London primary school, providing her with valuable 'frontline' experience of the challenges of urban education.*

*Prof John Stillwell is Professor of Migration and Regional Development in the School of Geography at the University of Leeds, Director of the Centre for Interaction Data Estimation and Research (CIDER) and Coordinator of the ESRC 'Understanding Population Trends and Processes' (UPTAP) initiative. He has worked with Education Leeds on using the PLASC data for analyses of pupil achievement, commuting and residential migration.*

*Dr Linda See is a Senior Lecturer in GIS and Computational Geography. Her research interests include soft computing techniques (fuzzy logic, neural networks and genetic algorithms) and agent-based modelling as applied to a range of spatial problems: flood forecasting, geodemographics, crime reduction, etc.*

# GIS enhances collaboration: using the line to draw disciplines together

Catherine Emma Jones[13], Laura Vaughan[2] Muki Haklay[1], and Sam Griffiths[2]

[1]Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT
Tel. +44 (0)20 7679 4188 | Fax + 44 (0)207 9161887
Email: ucfncej@ucl.ac.uk, www.sstc.ucl.ac.uk

[2]The Bartlett School of Graduate Studies, University College London, 1-19 Torrington Place, London, WC1E 7HB
Tel. +44 (0)20 679 1981 | Fax + 44 (0)207 9161887

[3]Department of Geography, University of Portsmouth, Portsmouth

KEYWORDS: Coordinate Line Geometry, Interdisciplinary GIS, Space Syntax, Spatial Analysis

## 1. Introduction

For many years, the rhetoric surrounding planning decisions in town centres has been driven primarily by retail activity. It is a commonly held belief that retail activity is the primary generator of the vitality and viability of high streets. In reality the composition of activities of town centres and their hinterlands are much more intricately interwoven. The research described in this paper was undertaken by the Towards Successful Suburban Town Centres Project (SSTC) at University College London. Whereby the project seeks to profile centres according to their socio-economic activity and morphological characteristics by drawing on a range of methods for the spatial analysis of social and economic activities at various scales. In this project the analysis of urban form using spaces syntax methodology is of particular interest. Indeed, initial analysis, conducted as part of the project, found that suburban town centres are complex, distinct centres comprised of multi-scaled interactions between land use activities, population and its built environment. This complexity of interactions contributes to the vitality of such centres as places for living, working, shopping and leisure (Vaughan et al, 2009).

To explore the detailed multi-scaled interactions of these centres, it is necessary to draw upon a range of methods to enable the geographical analysis of social and economic activities at various scales of accessibility and integration within the built environment. Of particular importance is the notion that network properties of the street plan exert an influence on pedestrian and vehicular movement patterns which have a bearing on the distribution of land use activities beyond just that of retail activity. An exploration of these relationships informs policy and planning decision making as well as urban design.

The type of geographic enquiry required to further our understanding lends itself to the examination of the complex relationships between land use activity and the structure of the built environment – bringing together interdisciplinary methods found in geography and architecture. The visualisation and quantification of (sub) urban form using space syntax methodology is of particular interest. It enables the spatial structure, accessibility and connectivity of streets and the layout of the buildings to be adjoined and then compared with information about the land use activities taking place within the suburban centres and its hinterland. To understand how the entities interact together, first they must be understood and only then can we model the relationships of how they interact.

The paper describes how space syntax methods developed in the field of architectural research to produce quantitative descriptions of built form and street networks can be enriched by using GIS as the mechanism for bridging between disciplines. The research illustrates how the resulting structural knowledge of the built environment can complement socio-economic analysis by providing a robust geographic base for measurement and analysis. This paper describes a methodology for disciplinary integration; first through the development of a point-line algorithm used to define and model the disjoint geographical relationship of entities in suburban town centres; and then how the resulting amalgamation between the GIS data alongside space syntax analysis acts as a collaborative tool for interdisciplinary investigation. It outlines the development of a cohesive geographical framework to aid the linkages between disciplinary practices and stresses the usefulness of the *line* as a predictor of movement potentials and as a suitable analytical measure for exploring land use activities.

## 2. Methodology - Building a collaborative model of suburban town centres

Research into the built environment has suffered due to a lack of collaboration and the incompatible nature of analytical toolsets by built environment specialists and human geographers (Jones et al., 2009). Space syntax theory and methodology is a powerful means of exploring the spatial configuration of the built environment. It uses quantitative analysis derived from graph theory and is represented by network maps (Hillier and Hanson 1984; Hillier and Iida 2005). Whereby the spatial structure of streets and the layout of the buildings are represented and analysed as a continuous spatial network in order to measure how well connected each street space is to its surroundings. This is done by taking an accurate plan of a built up area and drawing the set of least and fewest lines that cover all the open space ensuring that lines intersect where adjacent spaces are contiguous. Space syntax analysis then computes all the lines in the network according to their relative depth from each other. Depth increases with the number of changes of direction between lines, see Hillier, B. (2007). However, for human geographers this emphasis on physical structure seemingly marginalises socio-economic processes (Soja 2001), as it seems to imply a environmental determinism associated with the discredited positivism of geography's 'quantitative revolution' of the post-war period (Johnston and Sidaway 2004).

The geographers' critique is represented in figure 1(a) in which an over-determined built environment renders human activity as anonymous and mechanistic. By contrast, research in the field of space syntax argues that human geography prioritises the social construction of space at the cost of considering how this interacts with the built environment of the lived space (Hillier 2008). The world view of the human geographer, from the space syntax perspective, is illustrated in figure 1 (b) where human activity is represented against an undifferentiated background. Clearly there is a need to be sensitive to both perspectives if we are to understand the *relation* of the structure of the built environment to human activity, illustrated in figure 1 (c).

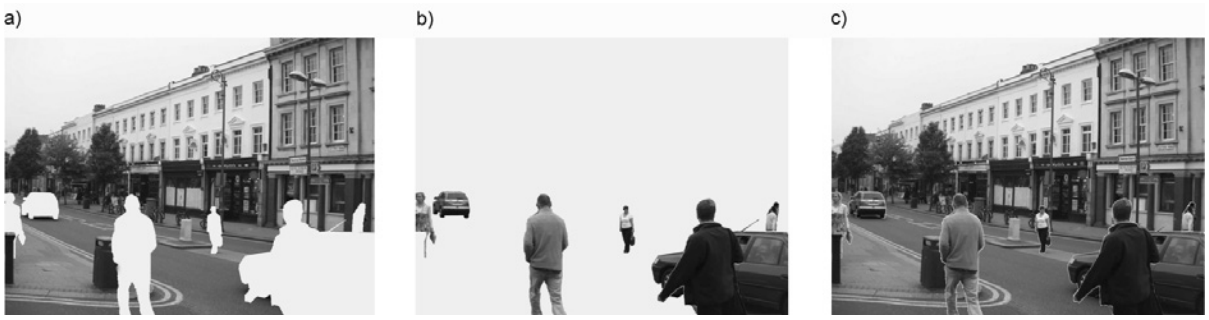



**Figure 1 (a-c):** Schematic illustration of contrasting approaches to society and space, from left to right: (a) built environment focused (b) human activity focused and (c) integrated

On the *Towards Successful Suburban Town Centres Project* it was sensible to consider this dialogue between researchers in human geography and space syntax in terms of the role GIS has in fulfilling its potential as an accessible facilitator of interdisciplinary research. The collaborative approach to using GIS technology therefore serves not only to highlight the competing philosophical approaches and methods of human geography and space syntax but also creates the possibility of a broad cross-disciplinary domain where these apparently competing ideas can be resolved. In order to unite the disciples, it is first necessary to identify two issues: (1) a lack of automated procedures that allow different graphical entities to be linked together[1] and (2) the absence of coherent shared methodologies and practices between the disciplines.

To develop interdisciplinary working practices, we first had to build a collaborative analytical framework with a geographical base to connect to the disparate entities; the syntax graph and the land use points (See figure 2). This is because they have dissimilar spatial forms, corresponding to the very distinct geographic components they characterise. The land uses are stored as points (figure 2b) and an abstract network of lines represents the shape and structure of the suburban built environment (figure 2c). We set out to develop a robust and consistent tool for modelling the relationship between the structure of built environment and functional land use distribution in suburban town centres. An algorithm based on coordinate geometry was written; its purpose, to define the correct spatial relationship between the lines and the points.

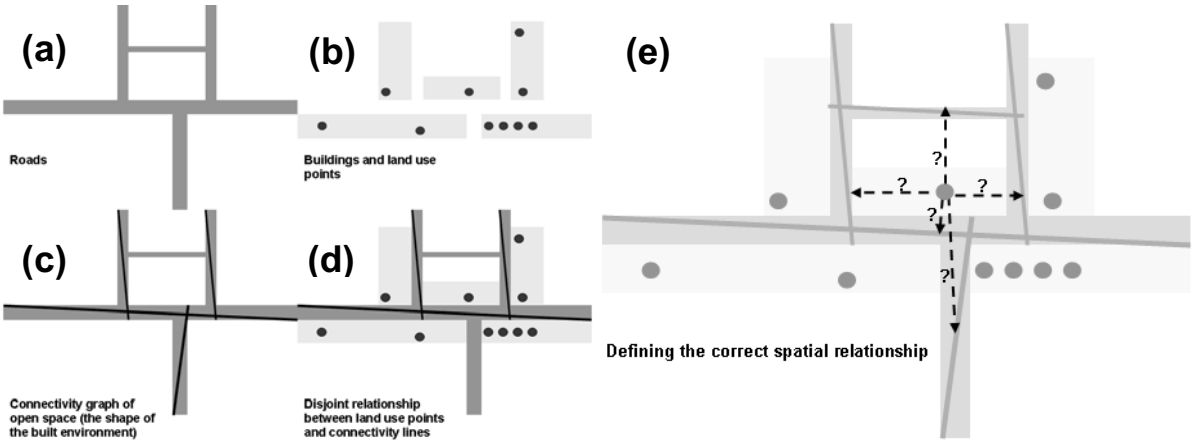



**Figure 2 a-e**: Building a cohesive geographic base for interdisciplinary analysis

Prior to modelling the geographical relationship of the objects, it was necessary to formalise the nature of the disjoint between the objects. In geometric terms, the central issue of the collaboration is one of a disjointed topology. Whereby space syntax methodology uses network graphs which are fundamentally abstracted from the social and geographical environment, they are commonly drawn by hand and have no attribute data that associates them to any base data such as addresses. Likewise the land use data is abstract from the built environment landscape as there is no cohesive attribute to link the two datasets (figure 1). The initial step in the methodology was to establish the number of address points located on an individual segment and determine what "*located on*" means, in the sense of the geographical relationship between the two entities (Figure 2e). Then each land use point could be linked onto the network graphic according to the appropriate geographical relationship.

[1] It should be pointed out that GIS is already used regularly in the discipline of space syntax for thematic mapping, although the methods for automated integration of points and lines is a relatively new development.

A search algorithm was developed based on the conceptual framework outlined in figure 3. Roads through the town centres were used as the unique identifier and to provide the common geographic base. Each land use point, derived from the Ordnance Survey Address Layer 2 product, has a full UK address attribute. It could be joined to the roads easily through the street name. The network graph representing the space syntax lines was also linked up to the town centre road network. Despite not having any attribute data, buffers were drawn around existing roads and the proportion of each syntax segment falling into a road buffer was subsequently calculated. The largest syntax segment section in a buffer was then assigned the road name corresponding to the buffer it was located within. Then, using the road and its name as the geographic base, the two disparate datasets could be linked.

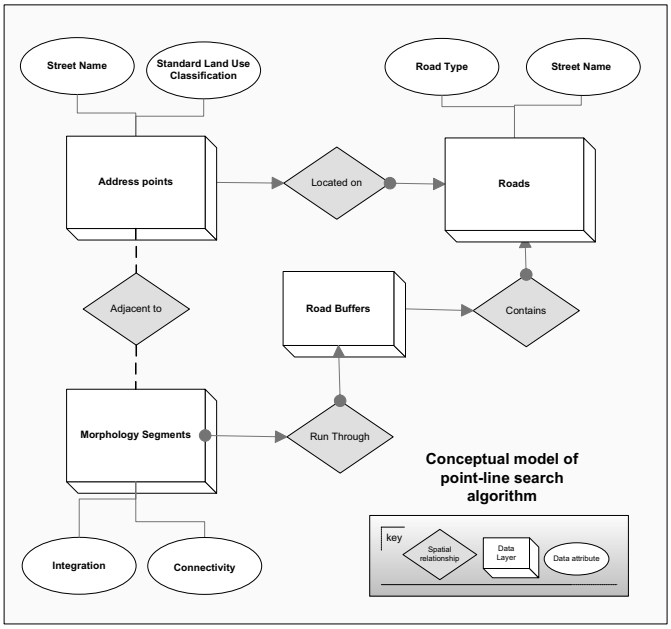



**Figure 3**: Conceptual model for defining a cohesive framework between land use activity and space syntax graphs.

## 3. Results and Conclusions

This new geographic base framework has great potential for interdisciplinary research. Initial findings have proven useful because the resultant data matrix was scrutinised further using analytical methods common to human geographers. The land use data associated to paths of movement potential in Town Centres underwent geographic investigations using Local Indicators of Spatial Autocorrelation analysis (LISA) (Anselin, 2006). The findings did reaffirm the geographical clustering of *retail* activity along highly accessible movement paths of high streets in town centres (see figure 1, Vaughan et al, 2009). More significantly, analysis of non residential activity and the potential movement of people in the built environment (Figure 4) demonstrates that routes through a town centre integrated at a scale of 1km are clustered significantly with all *non-residential* land uses activity. This reveals the importance of active segments beyond that of the retail High Street. It also suggests the geographical dependence of other types of non residential activity located off the high street and on locally integrated streets that are not routes where you would expect lots of through movement, indicating suburban high streets function not only as retail centres but that there is a more complex pattern of activity that ensures their vitality.

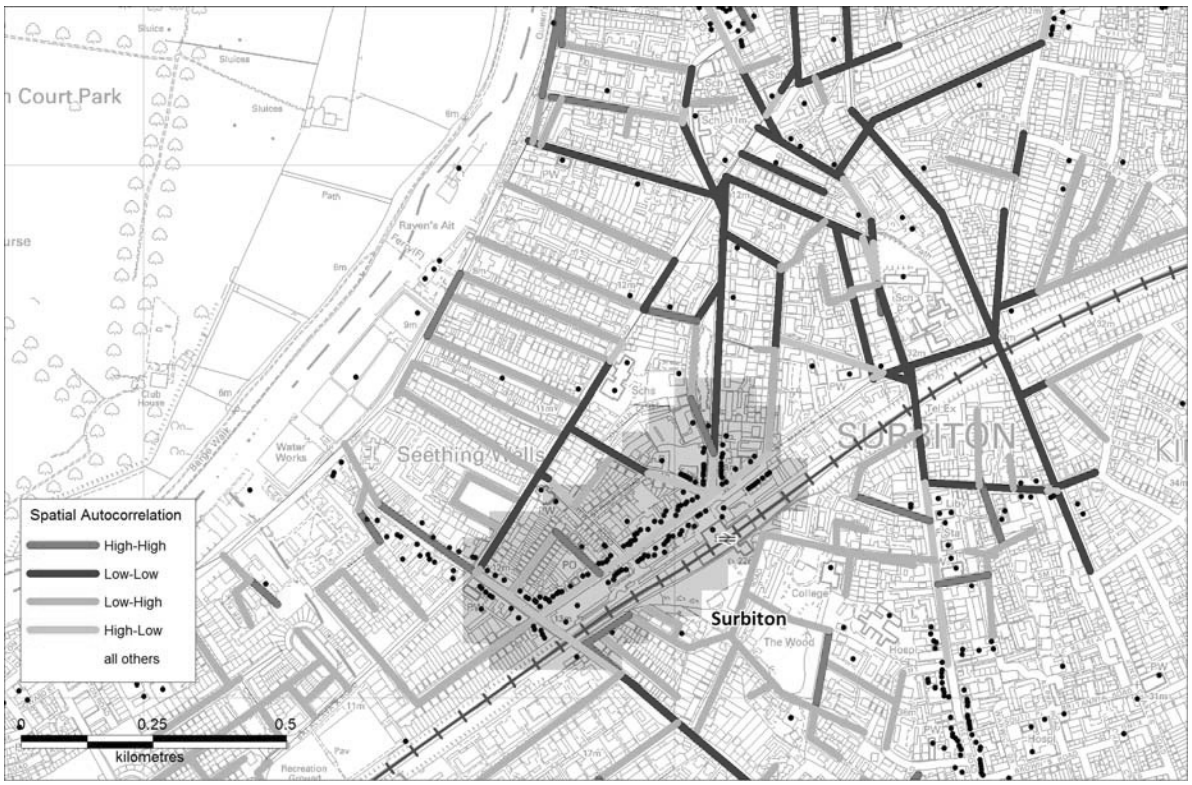



**Figure 4:** Significant clusters of segments integrated at a scale 1km in the town centre and the distribution of non-residential land use activities

The results of the algorithm led to the development of a coherent data structure based upon the network of movement potential in each town centre. The line entity became the object of priority and is symbolic of the linkages between disciplines. Thus, it is the semantic meaning of these two geographical objects that is of relevance to this paper. This fusion of the disciplines, the results of the geometric algorithm, represents the creation of a new phenomenon. The developed algorithm facilitated the amalgamation of two quite different world views with incongruent analytic frameworks. In one discipline the line object provided the mechanism for understanding the morphological properties of the town centres and in the other the point afforded investigation of the socio-economic properties of a centre's land use activities. Up until this point the two disciplines had not been as fully synthesised.

Furthermore linking the two disciplines produced findings with new insights that otherwise would not have been revealed. This synthesis between the disciplines also led to the development of a useful and usable analytical framework centred upon the line object. Thus, proving GIS as feasible mechanism for collaboration between Space Syntax and Human geography. GIS assisted in the development of automated procedures to allow the linking of different entities and the formulation of coherent shared methodologies and practices between the disciplines. The methodology developed in this paper contributes to the debate on the usefulness of GIS as a basis for driving collaborative research, illustrating how a geographical framework can be used to develop a methodology that is mutually beneficial for human geographers and space syntax specialists alike.

## 6. Acknowledgements

The research reported in this paper formed part of a 36 months study funded by the EPSRC (Engineering and Physical Sciences Research Council, start date 01/10/06. EPSRC reference EP/D06595X/1).

## References

Anselin, L., Syabri, L. I. and Y. Kho.Y. GeoDa: An Introduction to Spatial Data Analysis, *Geographical Analysis,* 38 (1), pages 5-22.

Hillier, B. and Hanson, J., 1984. *The Social Logic of Space*. Cambridge: Cambridge University Press.

Hillier, B., 2008. Space and Spatiality: What the Built Environment Needs from Social Theory, *Building Research and Information*, 36 (3), pages 216-230.

Hillier, B. and Iida, S., 2005. Network Effects and Psychological Effects: A Theory of Urban Movement, COSIT conference 2005 and published in A. Cohn and D. Mark (eds) *Spatial Information Theory Lecture Notes in Computer Science 3603*. Springer Verlag. pages 473-490.

Johnson, R.J. and Sidaway, J.D., 2004. G*eography and Geographers: Anglo American Geography since 1945*. London, New York: Oxford University Press.

Jones, C.E., Griffiths, S., Haklay, M. and Vaughan L. 2009. A multi-disciplinary perspective on the built environment: Space Syntax and Cartography - the communication challenge *Proceedings of the Seventh International Space Syntax Symposium*. Royal Institute of Technology, Stockholm.

Soja, E.W., 2001. In different spaces: interpreting the spatial organization of societies. *Proceedings 3rd International Space Syntax Symposium*. Atlanta, GA: 1, pages 1-1.7.

Vaughan, L., Jones, C.E., Griffiths, S. and Haklay, M. 2009. The Spatial Signature of Suburban 'Active' Centres. *Proceedings of the Seventh International Space Syntax Symposium*. KTH Royal Institute of Technology Stockholm.

## Biography

**Catherine Emma Jones**
Kate (Catherine) Jones completed her PhD in Geographies of Health in 2008. Having spent 2 years as a Post Doc at UCL she now has a lectureship in Human Geography at the University of Portsmouth. In the domain of health she is keen to peruse research into how the access, quality and visualisation of spatial data can be used for health intervention planning. She is interested in GIS as a toolset for interdisciplinary collaboration and geographic enquiry.

**Dr Laura Vaughan**
Laura Vaughan is Senior Lecturer in Urban and Suburban Settlement Patterns, UCL Bartlett School of Graduate Studies, She has written widely on urban and suburban settlement patterns, with a focus on space syntax as a method for fine grain analysis of built form. Currently she is Principal Investigator on an EPSRC grant - 'Towards Successful Suburban Town Centres: a study of the relationship between morphology, sociability, economics and accessibility'.

**Dr Muki Haklay**
Mordechai (Muki) Haklay is Senior Lecturer in Geographical Information Science at the Department of Civil, Environmental and Geomatic Engineering, UCL. He has been working on socio-economic analysis using GIS, novel analyses techniques with GIS and usability aspects of GIS. He has published papers on these subjects in Area, International Journal of Geographical Information Science, and the Journal of Environmental Management.

**Sam Griffiths**
Dr Sam Griffiths worked as a Research Fellow at the UCL Bartlett School of Graduate Studies 2006-9 and has since been appointed as a Lecturer. He has expertise in space syntax, urban morphology and historical methods.

# Exploring the Evolution of a Retailing System Using Visual Analytics and Simulation Gaming

Joel Dearden[1], Alan Wilson[2]

[1]Centre for Advanced Spatial Analysis, UCL, 1-19 Torrington Place, London, WC1E 7HB
Tel. +44 (0) 207 679 1782
Email j.dearden@ucl.ac.uk, www.casa.ucl.ac.uk
[2]a.g.wilson@ucl.ac.uk

KEYWORDS: urban modelling, simulation, visualisation, complexity, retail

## 1. Introduction

With over half the world's population now living in urban areas (Martine 2007) it is becoming increasingly important to understand how cities evolve so that we can ensure that urban environments are sustainable and serve the needs of their populace. Cities are systems of organised complexity (Weaver 1948; Jacobs 1961). The behaviour of any one component is dependent on the behaviour of many other components. This makes it difficult to analyse any one part in isolation. Allen and Sanglier (1981) highlight the difficulties associated with living in an interdependent society, where a change to any one part will have an effect on the rest of the city. Cities are also nonlinear systems which further complicates analysis. Collectively complex behaviours and structures such as segregation can emerge out of the micro-scale interactions of people (Mitchell 2009) but are not readily understood from their behaviour. The evolution of cities is influenced both by the actions of planners as well as the "self-organising" processes that occur in a city without any central control.

Dynamic urban models allow us to study and explore the process of urban evolution. Increases in available computing power have allowed for more detail to be represented by models but means that the models themselves are becoming complex and cannot be solved analytically - examples being agent based models (Schelling 1969) and Boltzmann-Lotka-Volterra models (Wilson 2008). Simulation is one way of analysing such models and involves running computational models and exploring the outputs. Such simulations provide a useful way of examining the underlying models and theory, identifying new and simple regularities and generating new hypotheses (Hartmann 1996). A simulation of this kind typically produces large volumes of data and involves complex relationships so we require good analytical tools in order to make progress. A visual analytics interface provides a powerful way to explore simulation outputs. The concept of scientific visualisation was introduced in the 1980s (McCormick et al 1987) as a way of bringing to bear the eye-brain system with its powerful pattern recognition capabilities. This is increasingly relied upon for analysis of large volumes of data where examination of the numerical data would be infeasible. Batty et al. (2004) introduce the idea of visual modelling and identify three potential benefits: (1) it allows simplification of complicated systems, (2) it enables exploration and discovery, and (3) it allows engagement with non-scientific experts. A recent extension to visualisation is the field of visual analytics, which can be defined as:

> *"…the formation of abstract visual metaphors in combination with a human information discourse (interaction) that enables detection of the expected and discovery of the unexpected within massive, dynamically changing information spaces" (Thomas and Wong 2004)*

Here the "dynamically changing information space" we are interested in is an urban system. By combining visual analytics with simulation we can allow a user to explore an urban model and the system represented by it in an immersive way and use their intuition to support analysis. An interface of this kind was demonstrated in the analysis of path dependence and discontinuities in an BLV urban

retail model (Wilson and Dearden forthcoming). Dearden and Wilson (forthcoming-b) derived the rules for a simple agent based model of urban retailing from the well known aggregate urban retail model (Harris and Wilson 1978) and used a visual analytics interface to develop the model and analyse the outputs.

Visual analytics software provides an interactive interface between a user and a computer simulation in which the user can view "real time" simulation outputs. From here it is a small step to give a user control of one or more agents and allow them to participate in the simulation. This simulation game could then allow the user to view, explore and experiment with the complex behaviour of an urban system. The system we envision differs from games produced by the entertainment industry in that the underlying mechanisms in the simulation are designed to be completely transparent to the user, the aim being to provide the user with an original viewpoint from which to critique the software and model and gain insight into the behaviour of the system of interest.

This paper presents preliminary work in developing a highly disaggregate and detailed model of urban retailing and using it as the basis for a simulation game that is set within a visual analytics framework. Potential users of such a system might include retailers looking to train staff, planners exploring the behaviour of a particular system and academics testing theories about retail systems. The complexity of the model means that the visual analytics and simulation game interface is essential for making sense of the simulation outputs as well as providing help in testing and debugging the model definition and simulation software. Such a system can help us to determine whether the combination of visual analytics and participatory simulation is useful for analysis of complex urban systems. We also aim to explore whether they can help overcome some of the criticisms of large scale urban models (Lee 1973).

We define the model in section 2, describe the visual analytics and user interface framework in section 3 and explain how a user will interact with the system in section 4 and conclude the paper in Section 5.

## 2. A disaggregate dynamic urban retail model

The simulation model described here is designed as an exploratory tool for generating hypotheses about the system and so aims be a stepping stone to producing better models of complex urban systems. Within the limitations of space, this paper can provide only a brief summary of the workings of the model. This is an agent based model in which the agents make probabilistic decisions. The main entities represented in the simulation are people acting as consumers and commercial companies acting as either property developers, retailers or both. Companies sell one type of good and can own both multiple shops and buildings allowing us to represent a range of company types from independent retailers through to large chain stores and property developers. Shops can vary in size from very small units right up to large supermarkets.

One iteration of the model is intended to represent one month of real time and moves through the following steps:

1. Identify clusters of widely spaced shops that make up "urban areas" (e.g. villages, towns, cities) and then, within each urban area we identify smaller "retail area" clusters of tightly spaced shops (e.g. high streets or shopping centres).
2. Consumers go shopping for each type of good they need - here we categorise companies and their shops into comparison and convenience goods. Consumer agents use a hierarchical decision making process to choose where to shop, following the steps: (1) choose an urban area, (2) choose a retail area within the chosen urban area, and (3) choose a shop within the chosen retail area.
3. Update company accounts – detailed accounts are kept for each company
4. Close bankrupt companies (those with a negative net worth) and replace them with a new start-up company. This results in the closure or sale of all their shops and buildings.

5. Determine strategy for each company in retail, property and borrowing. In each case the strategy can be one of: expand, contract or do nothing.
6. Allow each company to close unprofitable buildings and open new ones. New properties are either located randomly in the region or located by following a hierarchical decision making process based on revenue per unit area for all good types in each urban area, retail area and shop.
7. Allow each company to close unprofitable shops and open new shops. New shops are either located randomly in the region or located by following a hierarchical decision making process based on revenue per unit area for the same type of goods sold by the company making the decision in each urban area, retail area and shop.
8. Property values are calculated as a function of the revenue per unit area of the retail area the building belongs to. If the building is outside a retail area we set a minimum value.
9. We calculate shop operating costs based on premises rent, labour costs (assumed to be fixed based on a per unit area cost) and goods costs which are a function of the shop size and the size of the owning company (representing bulk buying economies of scale).

We initialise the model using real data for our test region which is the metropolitan county of South Yorkshire. All travel costs in the model are calculated in minutes using a road network constructed from Meridian 2 data - we calculate shortest paths between each entity in the simulation. The consumer agents are constructed from census data.

**3. Visual analytics and user interface**

A user interacts with the system by using an isometric three dimensional view of the region which can be customised with different views visualising particular variables using colour and shape, e.g.

- Shop gross income
- Shop net income
- Revenue per unit area of shops
- Ownership of each building and shop
- Consumer trip start and end points
- Shops shown by type of goods sold

Figure 1 shows an example view that displays shop locations and sizes (white blocks), urban area clusters (blue polygons) and retail area clusters (yellow polygons).

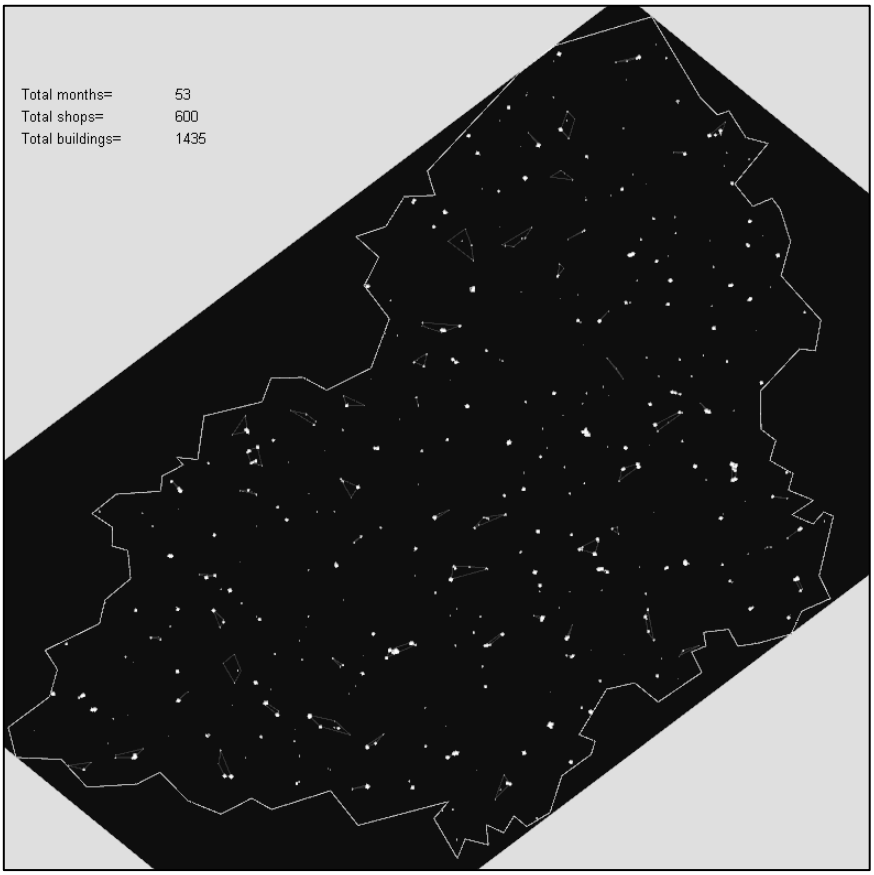



**Figure 1.** A simulation run showing the formation of urban and retail areas

## 4. Interaction

A user can take control of a single company within the simulation. The same rules that apply to agents constrain the player such as the need to maintain a non-negative net worth (total assets – total liabilities). Players are prompted to use the graphical user interface to make the following decisions during a model iteration, mirroring company agent behaviour:

- Decide which shops and buildings to keep open and which to sell
- Decide whether to open a new building. If they choose to open a new building they will need to decide its location and size (within the constraints of their budget).
- Decide whether to open a new shop. If they choose to open a new shop they will need to decide whether to buy or rent its premises. If they buy the premises they can choose the size and location. If they rent then they will need to choose from the available vacant properties on the market at that time.

If the user's company goes bankrupt they have the option of continuing the simulation in control of a new company.

## 5. Conclusions

The combination of visual analytics with participatory simulation has the potential to provide a useful and original point of view on the behaviour of dynamic urban models for the purposes of model development and improving our understanding of complex urban systems. By allowing a user to probe the model behaviour by direct interaction with the system we can enable better model testing and critique.

The model demonstrated here is very disaggregate and detailed. However this complexity appears to be manageable within an framework that uses object oriented programming, visual analytics and

simulation gaming. However the model contains so many variables that further work is required to test it ideally through the use of some kind of results grid, c.f. Dearden and Wilson (forthcoming-a). The model also needs to be properly validated against a known historical timeline, e.g. using the Town Centres data. Extensions to the system could provide support for multiple human users participating in one simulation over a network. The behaviour of company agents could also be refined to include the development of large scale planned shopping centres, mixed goods retailers and multi-purpose shopping trips.

**References**

Allen, P. M. and Sanglier, M. (1981), 'Urban Evolution, Self-Organization, and Decision Making', *Environment and Planning A*, *vol. 13*, no. 2, pp. 167-183.

Batty, M., Steadman, P. and Xie, Y. (2004), *Visualization in Spatial Modeling*, CASA Working Paper 79, Centre for Advanced Spatial Analysis (UCL), London, UK.

Dearden, J. and Wilson, A. G. (forthcoming-a), 'A Framework for Exploring Urban Retail Discontinuities', *Geographical Analysis*.

Dearden, J. and Wilson, A. G. (forthcoming-b), 'The Relationship of Dynamic Entropy Maximising and Agent Based Approaches in Urban Modelling', in *Spatial Agent Based Models: Principles, Concepts and Applications*, ed. A. Heppenstall, Crooks, A. and Batty, M, Springer, Berlin

Harris, B. and Wilson, A. G. (1978), 'Equilibrium Values and Dynamics of Attractiveness Terms in Production-Constrained Spatial-Interaction Models', *Environment and Planning A*, *vol. 10*, pp. 371-388.

Jacobs, J. (1961), *The Death and Life of Great American Cities*, Random House, New York.

Lee, D. (1973), 'Requiem for Large Scale Urban Models', *vol. 39*, no. 3, pp. 163-178.

Martine, G. (2007), *The State of the World Population 2007*, United Nations Population Fund, New York.

McCormick, B. H., DeFanti, T. A. and Brown, M. D. (1987), 'Visualization in Scientific Computing', *Computer Graphics*, *vol. 21*, no. 6.

Mitchell, M. (2009), *Complexity: A Guided Tour*, Oxford University Press, USA.

Schelling, T. C. (1969), 'Models of Segregation', *The American Economic Review*, *vol. 59*, no. 2, pp. 488-493.

Thomas, J. J. and Wong, P. C. (2004), 'Visual Analytics', *IEEE Computer Graphics and Applications*, *vol. 24*, no. 6, pp. 10–13.

Weaver, W. (1948), 'Science and Complexity', *American Scientist*, *vol. 36*, no. 4, pp. 536–544.

Wilson, A. G. (2008), 'Boltzmann, Lotka and Volterra and Spatial Structural Evolution: An Integrated Methodology for Some Dynamical Systems', *Journal of the Royal Society, Interface*, *vol. 5*, pp. 865-871.

Wilson, A. G. and Dearden, J. (forthcoming), 'Phase Transitions and Path Dependence in Urban Evolution', *Journal of Geographical systems*.

# Population 24/7: building space-time specific population surface models

Samantha Cockings, David Martin, Samuel Leung

School of Geography, University of Southampton, Southampton, SO17 1BJ, UK
Tel. +44 2380595519 | Email: s.cockings@soton.ac.uk

KEYWORDS: space-time, surface, population modelling, grid, spatio-temporal

## 1. Introduction

Many areas of social science research rely on small area representations of population. Current approaches to spatial population modelling rely almost exclusively on georeferencing of residential locations, drawing heavily on census definitions of 'resident population' and therefore essentially presenting an abstract representation of night-time population distribution (Bhaduri, 2008). There are however, good conceptual and practical arguments for modelling population at different times, incorporating population movements from seasonal to diurnal timescales so as to predict, for example, population exposure to a specific hazard or potential customer numbers during a working day. This paper presents early results from an ESRC-funded project to develop space-time specific population surface models of the UK. The project is based on an existing adaptive kernel density approach for building gridded surface population models (Martin, 1996), which is now being extended into a spatio-temporal kernel density estimation method. We begin by briefly reviewing relevant methods, then move on to our conceptual framework, data sources and modelling approach and conclude with some early illustrative results.

## 2. Space-time population modelling

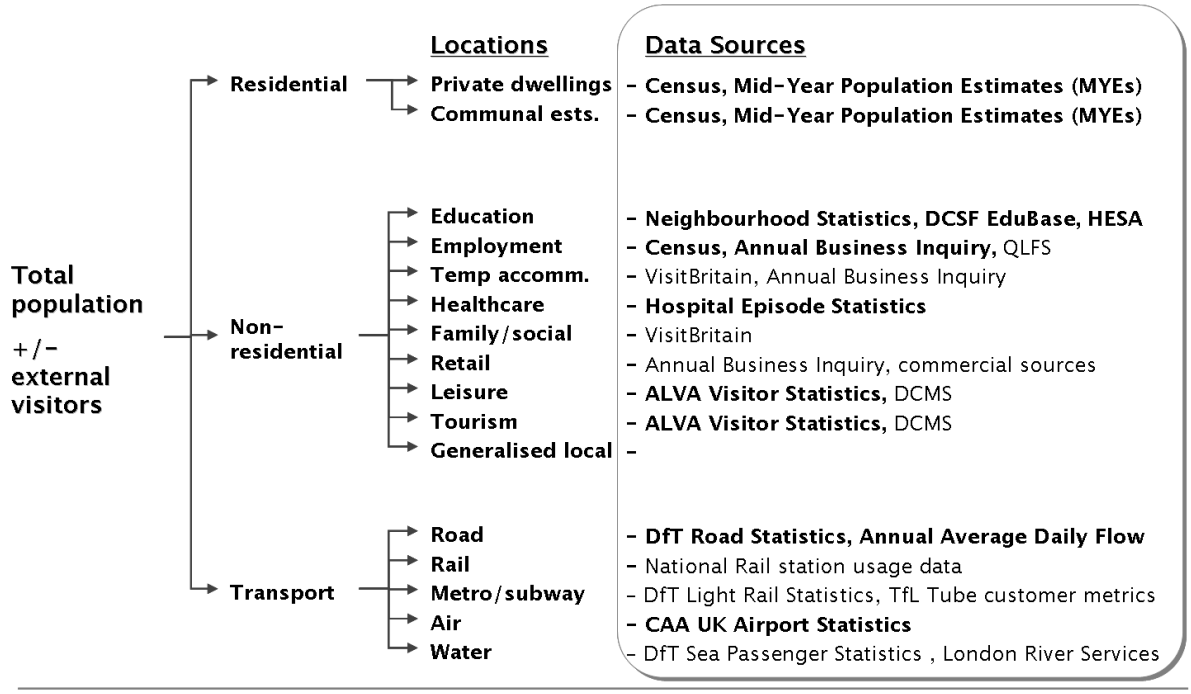
There are no widely accepted methods for obtaining time-specific population distribution models although the issue has long been recognised as important (e.g. Schmitt, 1956). Barriers to more fully-developed time-space population models have included both the weakness of GIS software for handling spatio-temporal phenomena and the absence of detailed data on short-term population movements.

Arguably, the most promising methodologies are based on dasymetric and/or gridded representations. Ahola et al. (2007) present a temporal model of Helsinki, using moving kernel density estimation for visualisation of results on a regular grid. The LandScan USA project (Bhaduri et al., 2007) is currently developing "daytime" and "night-time" 90m-resolution gridded models of the USA. Sleeter and Wood (2006) use US census data for small areas, transferring working and school populations from home areas during the daytime and redistributing these onto workplace and educational locations. McPherson et al. (2006) also attempt to model numbers of people in the transportation system. Smith and Fairburn (2008) provide perhaps the most relevant UK-based work, but their GIS database approach does not preserve population volume.

The temporal dimension is much more complex than just daytime or night-time modelling approaches. Overlaid on historical time are complex cyclical timescales including time of day, day of week, term times, public holidays and seasons. Population is further redistributed temporally and spatially on an ad hoc basis by special events. We aim to extend existing methods to incorporate these temporal complexities.

## 3. Conceptual framework

The underlying logic behind our conceptual framework is that each member of the population is engaged in one principal activity at one location at any one time. These activities and their locations can be successively partitioned, as in Figure 1, which is not exhaustive. Proceeding from the left, the entire population (adjusted for visitors) are either at residential locations, non-residential locations or in transit. Conventional census-based population mapping deals only with the residential category. Identification of high quality data sources for the other partitions would represent an enormous advance in the representation of population distribution in space-time, but even quite modest levels of subdivision would considerably advance current population mapping approaches. Data sources on the right hand box are indicative rather than comprehensive. So far within the project, we have used datasets shown in bold, others will be employed later.




Acronyms: **DCSF** Department for Children, Schools and Families; **HESA** Higher Education Statistics Agency; **QLFS** Quarterly Labour Force Survey; **DCMS** Department for Culture, Media and Sport; **ALVA** Association for Leading Visitor Attractions; **DfT** Department for Transport; **TfL** Transport for London; **CAA** Civil Aviation Authority

**Figure 1.** Population activities and potential data sources covering all or part of the UK.

## 4. Data sources

The method proposed here utilises existing census data, together with non-census data sources which have become available since 2001 and provide information about the location of population sub-groups at specific times. Smith and Fairburn (2008) provide a comprehensive review of relevant UK datasets. Many of these are collated within the Neighbourhood Statistics Service (NeSS). Others are published directly by government departments and other organizations, such as annual employer and employee counts; traffic flows; hospital inpatients and outpatients and tourist attraction visitors. By contrast, datasets from dynamic population monitoring (retail footfall, mobile telephone usage, vehicle tracking) are commercially sensitive, likely to be some years from realisation or present geoprivacy challenges (Nouwt, 2008). Even if all these sources were available now, there is no established integrative modelling framework.

## 5. Modelling framework

The surface modelling method developed by Martin (1989, 1996) and *SurfaceBuilder* software http://www.public.geog.soton.ac.uk/users/martind/davehome/software.htm, redistribute counts from population-weighted centroids of small areas into a regular grid using adaptive kernel estimation. Input centroids have thus far only represented residential locations and the resulting models have therefore been approximations to a conventional "night-time" population distribution on a reference date.

A spatio-temporal version of the model is now being developed to accept centroid locations, each of which is associated with a temporal profile describing the presence of population at different times. This combines the surface modelling approach with the spatio-temporal conceptualisation of Ahola et al. (2007). Standard time profiles (indicating the proportion of the capacity population that is present at any given time) are created e.g. for schools and different types of workplace or other activity. Seasonal, term-time, weekly and daily time cycles can all be incorporated within such a framework. Each model run relates to a specific target time and date, with the time profiles being used to weight the population at each centroid at the target time. Centroids are divided into "source" centroids (the sum of which defines the total population available for allocation) and "destination" centroids (to which population can be re-distributed at certain times). Each has a "region of influence" representing the extended area over which people may travel to that location. In some cases, these can be measured exactly (e.g. length of school day, catchment area), while others can be estimated (e.g. museum visitor numbers and origins from survey data). The temporal and spatial information interact – for example, a school will generate most travel in its area of influence at the beginning and end of the school day. There is no attempt to model individual activity patterns or transportation flows. Adjustments are, however, made for interregional and international passenger estimates applicable to the target time. The general approach is extensible to accommodate future data series and, importantly, retains the volume-preserving characteristics of the original algorithm.

## 6. Empirical example

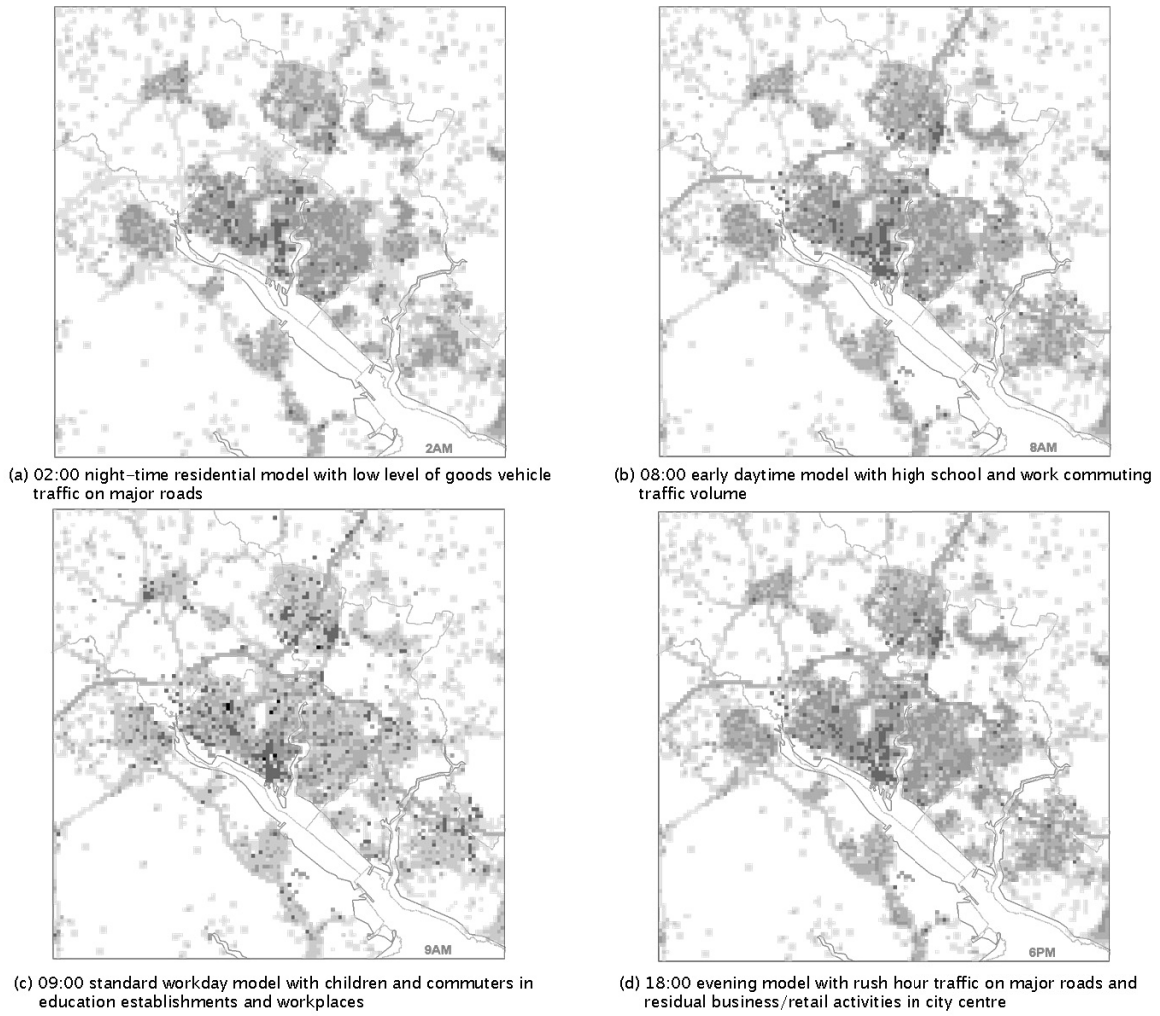
Our experimental work covers a Hampshire study area, including the City of Southampton. We have updated 2001 census-based counts to 2006 by allocating ONS MYEs proportionally to output areas (OAs) and unit postcodes. Our data processing is based on five age groups, determined by the available datasets. Census baseline information on 2001 workplace populations has been reconciled with Annual Business Inquiry (ABI) datasets at LSOA level and reweighted to OA level to obtain estimated workplace populations by broad industry type. Educational establishments from primary schools to universities, together with some age-sex breakdown of student numbers are available through NeSS, augmented with EduBase independent schools data to produce term-time population estimates for education locations.

Each of the aggregate datasets has been re-weighted onto centroid locations (usually OA or unit postcodes) if not available from the original source. Temporal profiles and areas of influence are estimated for each centroid type from a range of documentary sources. Some, such as school and college student numbers, opening hours, and catchment radii are relatively simple approximations, while time profiles of employment and distances of travel to work can be estimated from census and survey sources such as the national Quarterly Labour Force Survey. A wide range of activities such as visiting a neighbour, corner shop or family doctor fall below the spatio-temporal resolution of our modelling but do not involve the transfer of population out of residential areas. A "background" layer is being created which defines cells to which it is valid to allocate population counts, containing weights reflecting the capacity of the transportation network to contain population "in transit".

As a proof of concept and to validate the available datasets, we have used *SurfaceBuilder* to statically model the south Hampshire study population at different times, using the datasets highlighted in bold in Figure 1. Figure 2 illustrates early results, representing population densities for four reference times on a typical term-time weekday in 2006 at 200m resolution. Note that the figures do not yet contain correctly calibrated population in transit between locations. The four maps show the daily redistribution of population from an entirely residential (02:00) pattern; arrival of early workers at industrial and city centre workplaces (08:00), daytime distribution with major populations in education and office workplaces (09:00) and early evening when students and many workers have returned to residential locations (18:00).

## 7. Next Steps

The current stage of the project involves intensive software development, to be followed by the development of standard "runs" of the model for specified timescales. The model software and the standard outputs will eventually be available for download and use via the internet. We are also actively exploring the use of 3-dimensional block models and shaded polygon maps, overlaid in Google Earth (Figure 3) which, combined with time slider tools, can provide a powerful visualisation option for exploring time-space population distributions together with recognisable geographical features and placenames. The project website will be updated as work progresses: http://www.southampton.ac.uk/geography/research/rssa/pop247/index.html.




(a) 02:00 night-time residential model with low level of goods vehicle traffic on major roads

(b) 08:00 early daytime model with high school and work commuting traffic volume

(c) 09:00 standard workday model with children and commuters in education establishments and workplaces

(d) 18:00 evening model with rush hour traffic on major roads and residual business/retail activities in city centre

**Figure 2.** Gridded population models for south Hampshire study region on a term-time weekday in 2006 (25x25km, 200m cells). Dark shading: higher density. Coastline and MasterMap® Integrated Transport Network™ (ITN) layer data © Crown Copyright/database right 2010. An Ordnance Survey/EDINA supplied service.

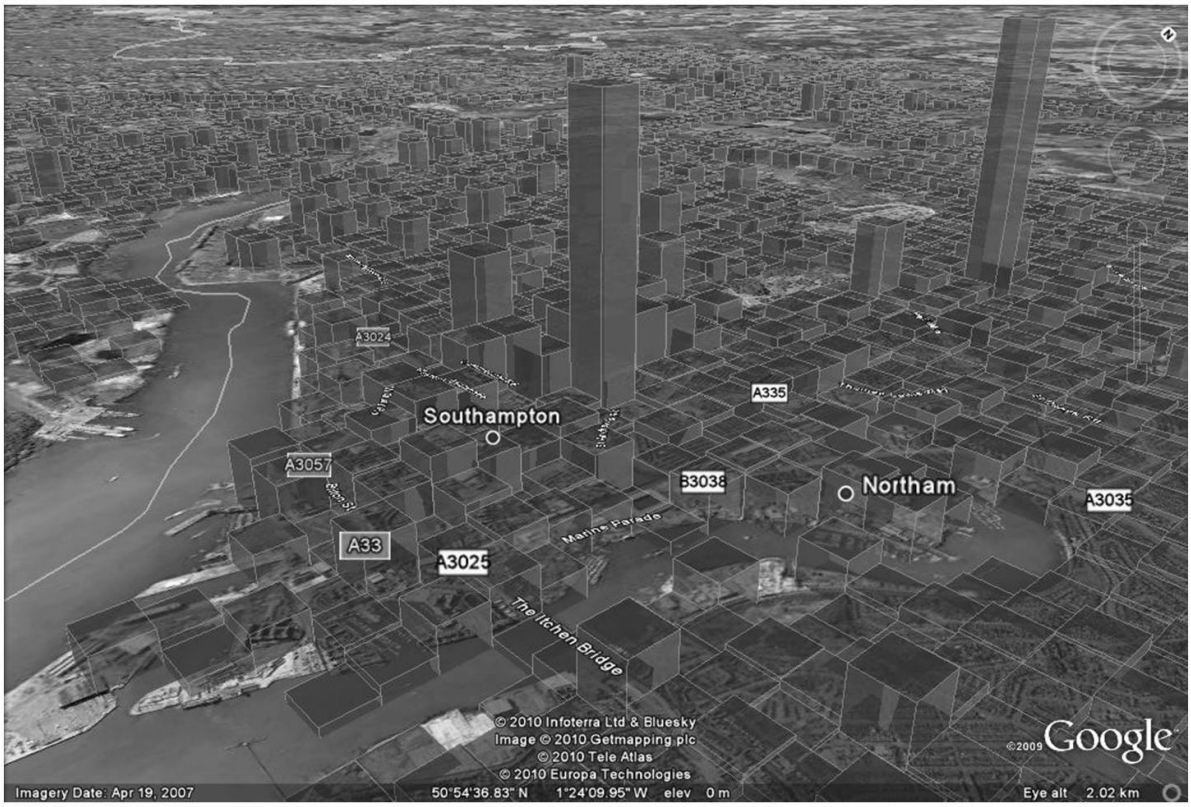



**Figure 3.** Daytime population densities in Southampton, rendered as a 3-dimensional KML layer in the Google Earth™ mapping service. © 2010 Tele Atlas © 2010 Infoterra Ltd & Bluesky © Europa Technologies Image © 2010 Getmapping plc.

## Acknowledgements

ESRC Award RES-062-23-0081. Annual Business Inquiry Service: NOMIS, licence NTC/ABI07-P3020. ONS 2001 Census: Standard Area Statistics (England and Wales): ESRC Census Programme, Census Dissemination Unit, Mimas (University of Manchester). National Statistics Postcode Directory Data: ONS, Postcode Directories: ESRC Census Programme, Census Geography Data Unit (UK-BORDERS), EDINA (University of Edinburgh). Quarterly Labour Force Survey: Economic and Social Data Service, usage number 40023. AADF Transport flow data: Department for Transport. EduBase: Department for Children, Schools and Families.

## References

Ahola,T., Virrantaus,K., Krisp,J.M. and Hunter,G.J. (2007) A spatio-temporal population model to support risk assessment and damage analysis for decision-making. *International Journal of GIS* 21, 935-953

Bhaduri,B. (2008) Population distribution during the day. In Encyclopedia of GIS Shashi Shekhar and Hui Xiong (Eds) 880-885 Springer

Bhaduri,N., Bright,E., Coleman,P. and Urban,M.L. (2007) LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *Geojournal* 69, 103–117

McPherson,T.N., Rush,J.F., Khalsa,H., Ivey,A. and Brown,M.J. (2006) A day-night population exchange model for better exposure and consequence management assessments. *Sixth Symposium on*

*the Urban Environment, Atlanta: American Meteorological Society*

Martin,D. (1989) Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers* 14(1), 90-97

Martin,D. (1996) An assessment of surface and zonal models of population. *International Journal of GIS* 10(8), 973-989

Nouwt,S. (2008) Reasonable Expectations of Geo-Privacy? *Scripted* 5, 375-403

Schmitt,R.C. (1956) Estimating Daytime Populations. *Journal of the American Planning Association* 22, 83-85

Sleeter,R. and Wood,N. (2006) Estimating daytime and nighttime population density for coastal communites in Oregon. *Proceedings of the Annual Conference of the Urban and Regional Information Systems Association*

Smith,G. and Fairburn,J. (2008) *Updating and improving the National Population Database to National Population Database 2*. Research Report 678, London: Health and Safety Executive

## Biography

*Samantha Cockings is a lecturer in Geography at the University of Southampton. Her research interests are focused on the socio-economic applications of GIS and in particular on the development of methodologies for the spatial representation of populations. Recent ESRC-funded projects include the Census2011Geog and Geo-Refer projects.*

*David Martin is a Professor of Geography at the University of Southampton. His research interests are focused on social science applications of geographical information systems. He is director of the ESRC/JISC Census Programme and co-director of the ESRC National Centre for Research Methods. He developed the SURPOP and SurfaceBuilder methodologies and software.*

*Samuel Leung is a research fellow with expertise in the handling of spatially referenced data and the development of online learning resources in geography. He is a land surveyor by background and previously taught at Oxford Brookes University before joining the University of Southampton to work on the JISC/NSF-funded DialogPLUS and ESRC-funded Geo-Refer projects.*

# Visualisation of UK Census interaction and housing market interaction data

Crispin Cooper

Department of City and Regional Planning, King Edward VII Avenue, Cardiff University
Tel. +44 (0)29 208 74022 | Email: cooperch@cf.ac.uk

KEYWORDS: visualisation, census, interaction, house prices, migration

## 1. Introduction

This paper presents two examples of the application of a technique useful for the visualisation of interaction matrices, allowing rapid, broad comprehension of the contents of complex datasets. The examples presented are UK Census migration data from the year 2000-2001, and inter-regional house price cross-correlations derived from Land Registry data, both at Local Authority level.

The traditional method of displaying such data would be with a flow map (Bertin 1984, page 350) although more modern techniques have also been used to calculate flow density (Rae 2009). In the first case, however, it is necessary to threshold the data (removing smaller flows of migrants) to ensure readability of the output, and both approaches miss patterns in the smaller flows of migrants which may be of interest to researchers.

By way of contrast, the pixel matrix plots presented here use data with a log function applied, which actually emphasizes smaller flows instead of trying to remove them from the plot. Pixel matrix plots lie in the tradition of Exploratory Data Analysis as advocated by Tukey (1977). While Bertin (1984) recommends displaying interactions in a matrix, and reordering matrix rows and columns to achieve greater readability, his methods of diagonalization and triangulation would produce different orderings based on the data values in different sets, all the while discarding spatial information. Instead, the approach taken is similar to Marble (1997) albeit with a more sophisticated technique for ordering matrix rows and columns (Guo & Gahegan 2006, Guo 2007), and the inclusion of pixel methods (Keim 1996, 2001).

Alternative approaches to visualisation of census flow data are discussed in Openshaw (1995), Kwan (2000) and Yan (2009), the latter of which uses a self organizing map to classify different types of interaction.

## 2. Methodology

An example of pixelation is shown in Figure 1. Cells of an interaction matrix representing flows between five locations, a-e, are shaded according to their values. It is then possible, without discarding information, to reduce the size of the graphic until each data point is represented by only one pixel.

The problem with this approach alone is that when a large number of origins and destinations are used, it is not easy to comprehend the plot because the X and Y axes don't represent anything real. It is better if a more intuitive ordering of the axes can be derived. The ordering aimed for is one in which places which are physically close together in 2-d space, are close together on the ordering; and *vice versa*. Thus, in the sense of Bertin (1984) the matrix itself becomes a *map*: a graphic where "the elements of a geographic component are arranged on a plane in the manner of their observed geographic order on the surface of the Earth" (page 285).

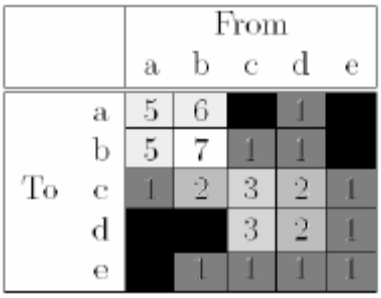



**Figure 1.** Pixelation applied to a simple interaction data set.

The method chosen to achieve the desired ordering is that proposed by Guo and Gahegan (2006), in which a variety of different algorithms are investigated for their ability to fulfil the criteria described above. The best of the algorithms investigated, CLO-OPT, was first developed in the field of bioinformatics (BarJoseph, 2003). In particular it outperforms the space filling curve techniques used by Marble (1997). The algorithm works by first hierarchically clustering all locations in geographical space, and then re-ordering the cluster tree to find the shortest path that visits all points. This has the effect that (i) urban areas formed by dense clusters of points tend to be kept together, and (ii) shortest path calculation is computationally feasible – $O(n^3)$, rather than O(n!) as is the general case without such a constraint.

A related re-ordering method is that of Wood (2009), which approximates a map whereby each originating region is itself replaced by a miniature map of the entire data set, showing the destinations of the flows originating from that region. A key difference from that method is symmetry: while Wood's method encourages contemplation of flows as properties either of their origins or destinations, the visualisations presented here treat both symmetrically, thus emphasizing the structure of the interaction matrix itself (albeit at the cost of a less intuitive pixel ordering). Thus, each method will have a tendency to emphasize different patterns in the data.

Interactive software has been developed to assist reading of the plots presented here; an example screenshot is shown in Figure 2. This also illustrates the ordering of UK Local Authorities chosen by the algorithm.

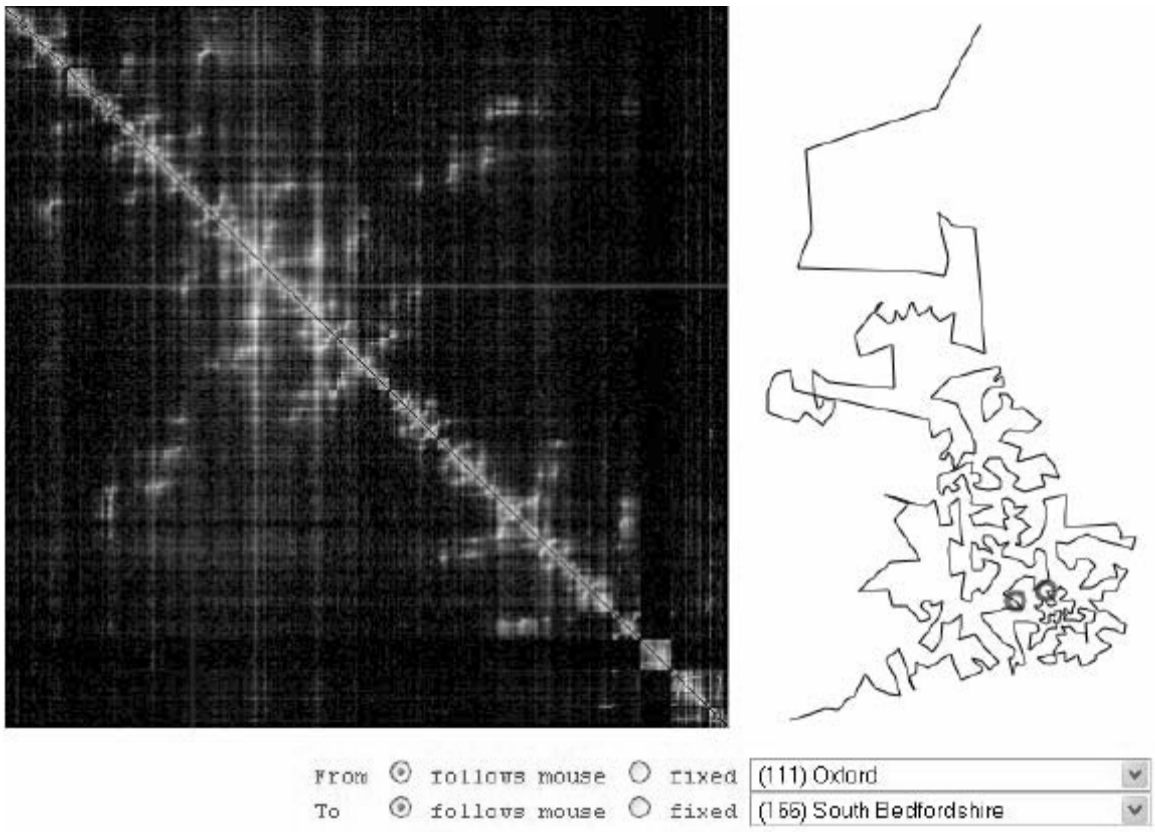


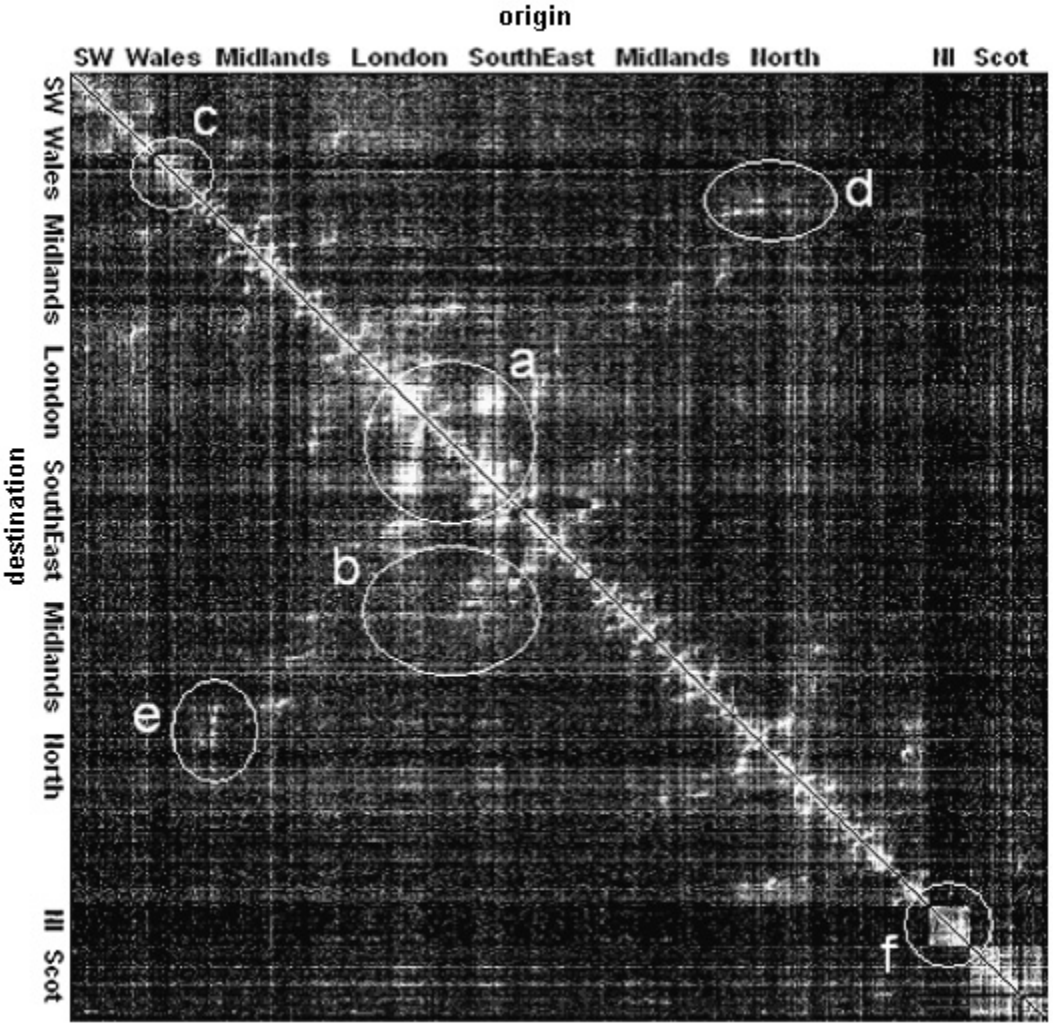

**Figure 2.** Screenshot of interactive software developed to assist exploration of pixel matrix plots. The map to the right hand side displays the linear ordering of points in 2-d space, with the current origin and destination highlighted on the map by red and green circles and on the matrix by lines of the corresponding colour. The data displayed are commuting flows from the 2001 Census.

### 3. Visualisation of intra-UK migration flows

Figure 3 shows the resulting visualisation of intra-UK migration. The ordering presented is the same as that shown in Figure 2. Several patterns in this plot are worthy of discussion.

- The grouping of the lightest pixels towards the diagonal of the image shows that the vast majority of migrations take place on a local basis.
- The fact that the remainder of the plot appears to consist mainly of vertical and horizontal lines, shows that for non-local migrations, distance is not a deciding factor; rather it is the inherent repulsiveness of origins and attractiveness of destinations that determine migration flows.
- The four yellow squares of feature **a** represent London, for which a lifecycle migration pattern is visible. Thick orange lines, extending horizontally outwards from region **a**, indicate a flow of younger people from all over the country migrating into the capital. Fainter green lines extending vertically out of region **a** indicate the middle-aged leaving London; while the strong blue patch below (marked **b**) shows an older population migrating from London to East Anglia and surrounding areas. The yellow colour of London itself shows inter-London migration to be dominated by 16-40 year olds.
- The City of London is clearly visible as a black cross centred in the lower right yellow square. This is because it has little residential population and therefore very few migrations to and from the City occur.

- Urban polycentricity is visible in London. This is as defined by Hall (2001) who notes that rather than having a single centre which exceeds all other parts of the region in its provision of products and services, London is "now the centre of a system of some 30-40 centres". The pixel matrix plot visualisation arguably shows that Greater London is polycentric in terms of migration movements, because no discernible internal structure is visible within the yellow squares of region **a**. This is in contrast to other areas in the UK, for example region **c** which represents South Wales, with the bright internal cross shape representing Cardiff – which is clearly a monocentric keystone of interaction for the region.
- The feature marked **f** represents Northern Ireland, easily identifiable because of its strong internal structure but having with little interaction with the rest of the UK.




| Flow volume | Age 16-25 | (mix) | Age 25-40 | (mix) | Age 40+ | All ages |
|---|---|---|---|---|---|---|
| 0 | | | | | | |
| ~1-50 | | | | | | |
| ~50-1000 | | | | | | |

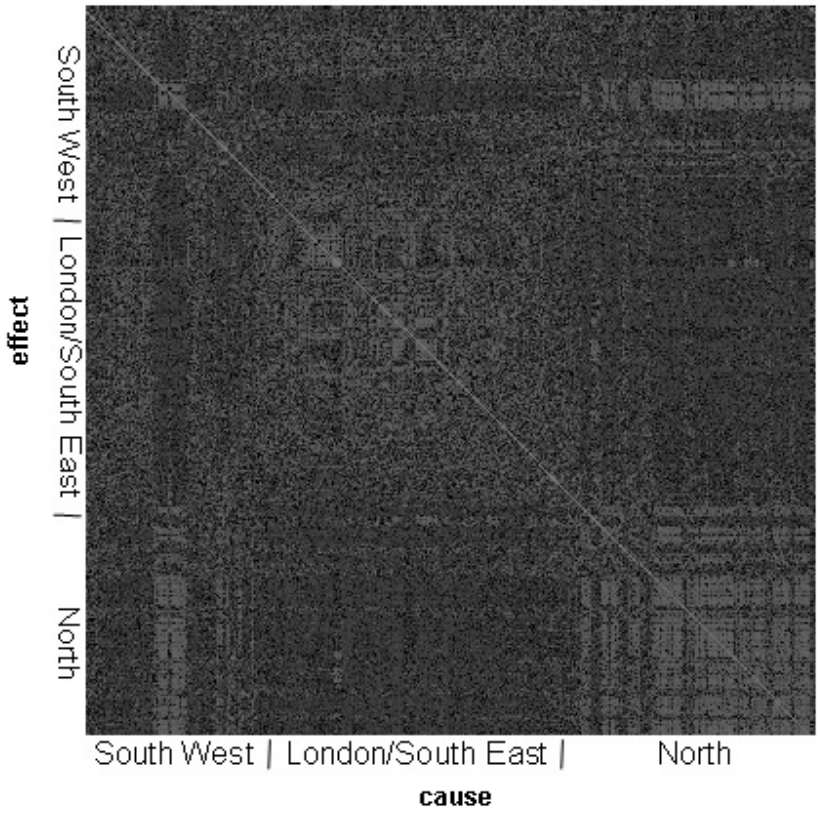
**Figure 3.** Pixel visualisation of UK Local Authority migration flows for the years 2000-2001, with logarithmic scaling. The features labelled a-f in white are discussed in Section 3.

**4. Visualisation of UK housing market cross-correlations**

Figure 4 shows a similar plot for cross-correlations in the housing market. Cross correlation for an

origin A and destination B is defined as the coefficient of correlation between house price increases at A, and price increases at B over the following 200 days. The pattern exhibited is remarkably different to the pattern of migration shown in Figure 3. Two key features of the structure of the housing market as presented in this plot are:

- Large blocks of red along the diagonal axis, indicating large areas with strongly correlated house price time series;
- Numerous horizontal and vertical lines of similar-coloured pixels, indicating the existence of certain places which tend to drive the market more strongly than others, or are driven by the market more strongly than others.




| Correlation | Colour |
|---|---|
| Greater than average | |
| Less than average | |

**Figure 4.** Pixel visualisation of England & Wales house price time series cross-correlations, 2000-2006.

## 5. Conclusions

It should be noted that the data features seen in these visualisations should not be taken as complete deductions, but as hypotheses for further (more rigorous) deductive testing. In the vein of Exploratory Data Analysis, the primary value of these plots is in hypothesis generation. It should also be noted that the patterns seen are heavily dependent on the order of pixels in the matrix; therefore the absence of a visible pattern should not imply the absence of a feature in the data.

A key limitation of the technique is that any linearization of 2-d space will necessarily not be a

perfect representation of that space. Thus, some features seen in the plot are artefacts of linearization rather than real features of the data. The features marked **d** and **e** in Figure 3 are an example of this. For comparison, Figure 5 shows a visualisation of inter-LA distances for England and Wales. It can be seen that the majority of points separated by <50km are close to the diagonal; however, a few problematic regions exist.

The primary justification of the plots presented here is that the author has found them useful in the analysis of large data sets, in the development of models of the data and in the debugging of related software. While such displays of information require a certain amount of practice to read effectively, this effort enables quick viewing of more patterns in the data than would be discernable by most existing techniques.

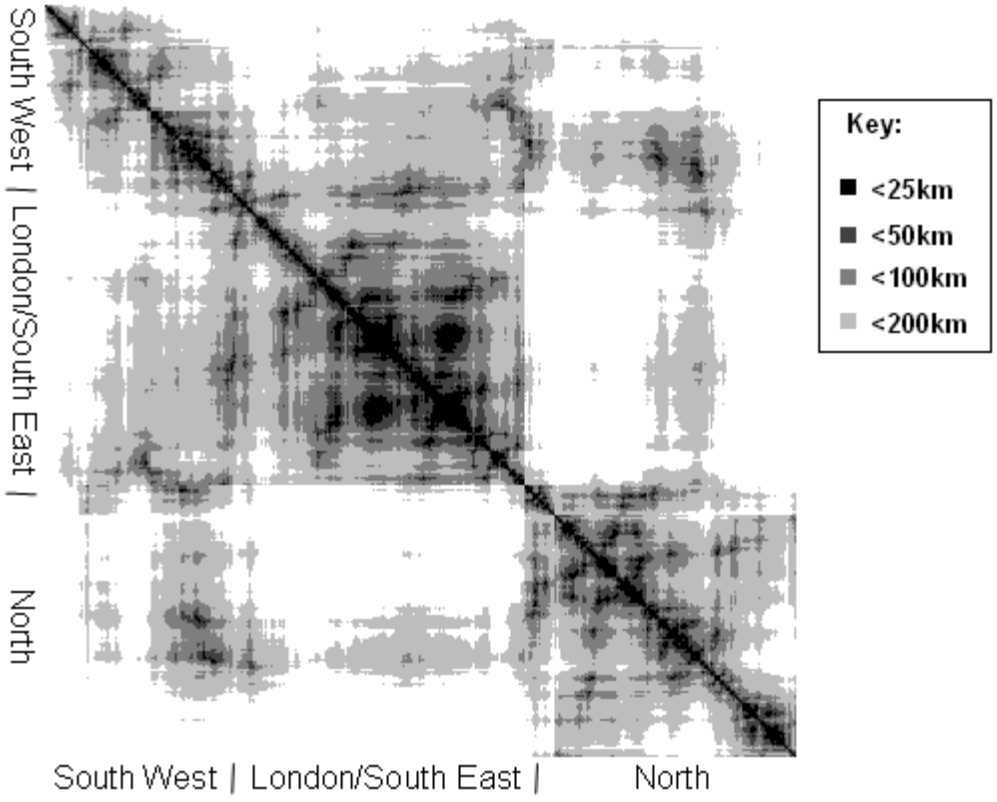



**Figure 5.** Visualisation of England/Wales inter-LA distance, illustrating the quantity and nature of linearization artefacts endemic to the technique.

## 6. Acknowledgements

The author would like to thank the Landmark Information Group ltd for providing the house price data used in this paper.

## References

Bar-Joseph et al (2003), K-ary clustering with optimal leaf ordering for gene expression data, *Bioinformatics* 19(9), 1070-1078'.

Bertin, J. (1984), *Semiology of Graphics*.

Guo, D. (2007), Visual analytics of spatial interaction patterns for pandemic decision support,

*International Journal of Geographical Information Science* 21(8).

Guo, D. & Gahegan, M. (2006), Spatial ordering and encoding for geographic data mining and visualisation, *Journal of Intelligent Information Systems* 27, 243-266.

Hall, P. (2001), Christaller for a global age: Redrawing the urban hierarchy, in *Stadt und Region: Dynamik von Lebenswelten, Tagungsbericht und wissenschaftliche Abhandlungen*, 53. Deutscher Geographentag Leipzig. URL: http://www.lboro.ac.uk/gawc/rb/rb59.html

Keim, D. A. (1996), Pixel-oriented database visualizations, *SIGMOD record* 25(4).

Keim, D., Hao, M. C., Ladisch, J., Hsu, M. & Dayal, U. (2001), Pixel bar charts: a new technique for visualizing large multi-attribute data sets without aggregation, in *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*.

Marble, D., Guo, Z., Liu, L. & Saunders, J. (1997), Recent advances in the exploratory analysis of interregional flows in space and time, in *Z. Kemp, ed.,Innovations in GIS 4,* Taylor and Francis.

Openshaw, S., ed. (1995), *Census Users's Handbook*.

Rae, A. (2009), From spatial interaction data to spatial interaction information? geovisualisation and spatial structures of migration from the 2001 census, *Computers, Environment and Urban Systems*. doi:10.1016/j.compenvurbsys.2009.01.007.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison Wesley.

Wood, J., Dykes, J., Slingsby, A. and Radburn, R. (2009), Flow trees for exploring spatial trajectories, in *Fairbairn, D., Ed., Proceedings of the GIS Research UK 17th Annual Conference*, pp. 229-234, University of Durham, Durham, UK.

Yan, J. & Thill, J.-C. (2009), Visual data mining in spatial interaction analysis with self-organizing maps, *Environment and Planning B* 36.

**Biography**

*Crispin Cooper began his research career in the Intelligent Systems group at the University of York. He is now completing his thesis on advanced numerical analysis of census and house price data at the department of City and Regional Planning, Cardiff. His first degree is in Computer Science (Cambridge, 2002).*

# Analysing Uncertainty in Home Location Information in a Large Volunteered Geographic Information Database

Nazanin Khalili, Jo Wood, Jason Dykes

giCentre, School of Informatics, City University London EC1V 0HB
Tel.+44 (0)20 7040 0146 | Fax +44 (0)20 7040 8584
Email: nazanin.khalili-shavarini.1 | jwo | jad7 @soi.city.ac.uk

KEYWORDS: vernacular geography, vagueness, VGI, spatio-social relation, *Flickr*

## 1. Introduction

This paper examines the ambiguity and location uncertainty in volunteered geographic information or VGI (Goodchild, 2007). Considering the multi geographies associated with individuals through VGI, the naive geography of confining social entities in a bounding box of a city is not adequate for study and analysis of their complex spatio-social relations. *Flickr* (Yahoo's photo sharing site) is a rich source of spatio-social data among a spatially structured social group (Khalili et al, 2009). The unstructured *flickr* Hometown Location Information (FHLI) associates individuals with places and varies from vernacular geographical terms to precise coordinates. We introduce a method for classifying and disambiguating the uncertainty in FHLI geography to augment bounding box geography and support geographical analysis of FHLI data.

## 2. Previous Work

There is growing body of work in disambiguating vague, indeterminate location information (Amitay et al. 2004; McCurely, 2001; Jones et al. 2008; Jacquez et al. 2000; Montello et al. 2003). Existing solutions either describe place names according to their structure, the context in which they are examined or refer to gazetteers that are used with GIS. While the spatial terms used in GIS are assigned sharp boundaries, the user generated geographic information on the web is inherently imprecise and fuzzy (Silva et al, 2006). Consequently, the existing disambiguation methods are vulnerable when applied to VGI as:

- They are only applicable to structured databases (Amitay et al, 2004)
- Location information in VGI does not conform to existing geographic boundaries (Purves et al. 2009; Jones et al. 2008)
- Gazetteers exclude vague terms (Popescu et al. 2008).

## 3. Flickr

Since *flickr* does not apply any restrictions on how users can define their home locations, FHLI varies widely from informal natural language (vernacular geography) to more formal scientific geographical vocabulary and numeric information. The analysis of what Waters and Evans (2003) term "fuzzy psychogeographical" can reflect the people's behaviours in defining geographic places on the web. However, its multifarious nature means that such work is by no means straightforward.

### 3.1 Sample Dataset

In order to produce an unbiased but indicative dataset that is rich enough for further analysis we generated a structured sample of FHLI data. In consideration of representative and

manageable dataset for better analysis and ease of mapping fifteen photos of the highest available accuracy were randomly selected on a daily basis (Table1). Since the sampling periods cover entire *flickr* life from its initial launch there is potential for analysis of changes in geotagging behaviour and friendship network over time.

**Table 1.** Number of posters for the randomly selected photos with their unique home locations.

| Intervals | poster | Unique Home Location |
|---|---|---|
| 19/07/08-18/07/09 | 1,142 | 584 |
| 19/07/07-18/07/08 | 991 | 461 |
| 19/07/06-18/07/07 | 1043 | 481 |
| 19/07/05-18/07/06 | 947 | 426 |
| 19/07/04-18/07/05 | 846 | 381 |
| 01/02/04-18/07/04 | 411 | 192 |

### 3.2 Specification

Our initial attempts to classify the FHLI revealed six classes of terms that required new disambiguation methods to be developed and new precision measurements to be applied (Table 2). We order these according to different types of vague terms occurred in the dataset.

**Table 2.** Classes of vague terms in FHLI.

| Vague Classes | Examples |
|---|---|
| **Doesn't exist** | • never: '*Outer space', 'L???*'<br>• current: '*Standel, Kent County*' |
| **Multiple Alternatives** | • same name for different places<br>  ○ multiple scale: '*Netherland* (country, city, town, hamlet)<br>  ○ multi-site place: '*Nanyang technological University'*<br>• different names for same places: '*Germany, Allemand, Deutshland*' |
| **Multiple Entities** | • *'UK, Paris one day Italy'* |
| **Abbreviation** | • single scale: *'Philly' (Philadelphia)*<br>• multiple scale: *'PC, US'(Pacific Coast, Panama City, Park City, Penn Central)* |
| **Mis-spelling** | • 'Toru?, Poland' |
| **Descriptive** | • 'I live somewhere with lots of sunshine' |

## 4. Methods

In the light of the above cases, in order to study how people define their home locations on the web a method is required that can successfully complete the following three steps:

1. Disambiguation
2. Precision Measurement
3. Uncertainty Classification

The majority of the existing algorithms for disambiguating the vague terms are based on very strong discourse effects between words in a single document. Therefore, they apply the discourse constraints through probability (Smith and Mann, 2003) or one sense per discourse analysis (Gale et al. 1992; Li et al. 2003). As the nature of the FHLI is not that of a well edited single document, these methods cannot independently disambiguate the home locations successfully. We have therefore, applied and adopted both methods according to FHLI specifications (section 3.2).

The method we propose considers:

***P*** (London, London UK) > ***P*** (London, London Ontario) If
***Occurrence*** (London UK) > ***Occurrence*** (London Ontario)

(P stands for probability)

In cases in which the occurrences of the alternative places are of equal value or there is no occurrence of the alternative names in the dataset, the disambiguation is achieved by selecting the case with higher population. This approach, adopted from Rauch et al. (2003) itself relies upon an uncertain entity that may be associated with an uncertain extent.

The next step is to measure the precision of the disambiguated names. In the first step attempts were made to apply the *flickr* geo photos' classification model:

- World level = 1
- Country ~3
- Region ~6
- City ~11
- Street ~16

Considering the fuzzy vernacular geographic terms that are frequently found in FHLI and in order to achieve a more spatially precise classification, we have extended the *flickr* model to include more detailed hierarchical spatial units. Accordingly, fourteen distinct precision levels were identified (Table 3).

**Table 3.** Precision classification for FHLI.

| *Precision Level* | *Spatial Unit* | *Precision Level* | *Spatial Unit* | *Precision Level* | *Spatial Unit* |
|---|---|---|---|---|---|
| 0 | Blank | 5 | Region | 10 | Village |
| 1 | Unknown | 6 | State | 11 | Street |
| 2 | World | 7 | City | 12 | Postcode |
| 3 | Continent | 8 | Town | 13 | House No. |
| 4 | Country | 9 | Borough | 14 | Coordinates |

Classifying the home locations (section 3.1) according to spatial units identified above have resulted in some inconsistencies which were due to the facts summarized in Table 4.

**Table 4.** Inconsistencies in measuring precision for the spatial units.

| Description | Example |
|---|---|
| Different internal administrative names for land units in each country | *'Parroquia'* in Spain, '*Ward'* in Japan. |
| Different internal organizations (land divisions) exclusive to each country | *'Province'* in China and Canada. |
| Inconsistency between size and population and the hierarchy of administrative divisions. | *'Ipswich'* (town) larger than *'St Davids'* (city) within a single country – Britain. |

Accordingly, in order to minimize the mentioned inconsistencies in spatial units across nations the population of each alternative spatial unit is also considered for the precision measurements of FHLI. A suitable uncertainty number (from 1 to 5) reflects our confidence in this classification in each case (Table 5).

**Table 5.** Uncertainty classification for FHLI

| Uncertainty Classification | Description |
|---|---|
| 1 | Less uncertain than the following uncertainties*('London, UK'*) |
| 2 | Nested spatial units e.g. city and county *('Denver', 'New York'*) |
| 3 | Different places in one country*('Portland, US', 'Cangas, Spain'*) |
| 4 | Different places in different countries *('Netherland'*) or several places for a single user ( '*Anchorage, Los Angeles, Someday New York, may be Paris'*). |
| 5 | Blank or information that cannot be associated with any place in the world ('*Outer Space, L????'*). |

## 5. Results

Classifying the sampled FHLI according to the proposed method can demonstrate how precisely people refer to their home locations on the web. As Figure (1) demonstrates there are remarkably consistent patterns in all the examined time periods with the most significant number for unknown and city level.

Figure (2) illustrates the percentage of FHLI in each category associated with each of the five types of examined uncertainties. Comparing the UC3 to UC4 (Table 5 and Figure 2) indicates that uncertainty and ambiguity in place names are more considerable in national level than across nations.

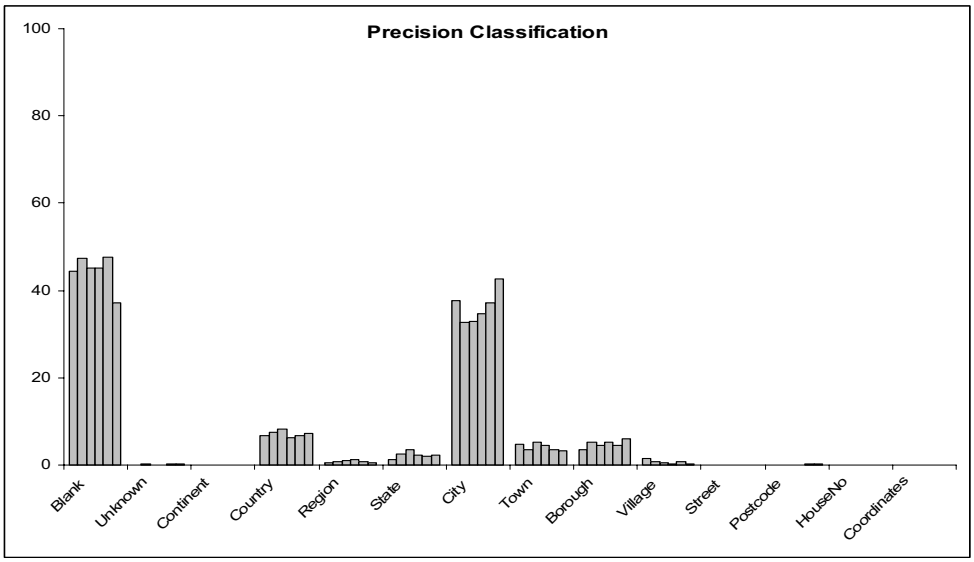



**Figure 1**. Precision classification for the sample data

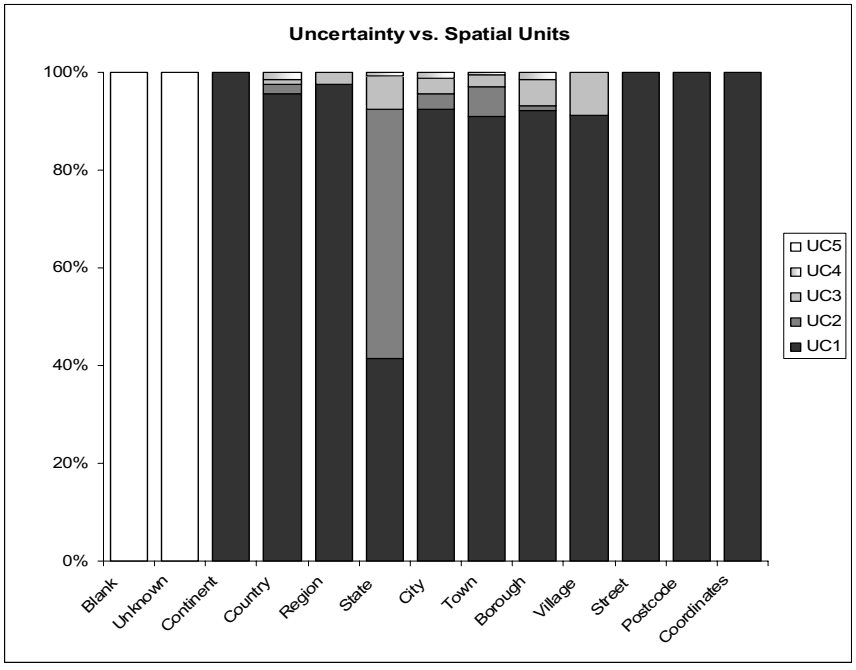



**Figure 2.** Uncertainty assigned to spatial units.

## 6. Conclusion

This paper examines the complexity in classification and disambiguation of the vague geographic terms in VGI. The above preliminary analysis is conducted in order to assess and evaluate the specifications of the FHLI. The initial results are expected to contribute towards selecting a rich, unbiased and indicative sample FHLI. According to the introduced classes of vague terms in FHLI (Table2) and the fuzzy vernacular nature of the dataset, it is not feasible to fully automate the proposed approach. However, there is potential to automate the

classification process for some of the FHLI by referring to the location classifications used in gazetteers (Smith and Mann, 2003; Hill, 2000). Overall, this paper concludes that:

- Focusing on the national scope of FHLIs can increase the ambiguity and uncertainty in FHLI and
- Analysis of the distribution of geotagged photo collections in line with analysis of FHLI might improve our confidence in classification and disambiguation process.

Therefore, we plan to apply this method to a broader subset of geotagged photos relating to the UK during the same time periods. Geotagged photo collections will be used as extra location information for estimating home locations. The assigned uncertainties to the disambiguated terms will be used in analysis of the FHLI and visualization of the multi geographies associated with social entities. The final models and techniques are expected to be applicable to variety of VGI available on the web.

## References

Amitay, E., Har'El, N., Sivan, R. and Soffer, A. (2004) Web-a-Where: Geotagging Web Content. 27th annual international conference Special Interest Group on Information Retrieval (ACM SIGIR), Sheffield, UK, 273-280.

Gale, W.A., Church, K.W. and Yarowsky, D. (1992). One Sense Per Discourse. Proceedings of the 4th Defence Advanced Research Projects Agency (DARPA) Speech and Natural Language Workshop, 233-237.

Goodchild, M.F. (2007) Citizens as sensors: the world of volunteered geography. GeoJournal, 69, 211-221.

Hill, L.L**.,** (2000) Core Elements of Digital Gazetteers: Placenames,Categories, and Footprints. In Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, J.L.Borbinha and T.Baker, Eds. Lceture Notes in Computer Science, 1923, 280-290.

Jones, C., Purves, R.S., Clugh, P.D. and Joho, H. (2008) Modelling Vague Places with Knowledge from the Web. International Journal of Geographical Information Science, 22(10), 1045-1065.

Jacquez, G., Maruca, S. and Fortin, M. (2000) From Fields to Objects: A Review of Geographic Boundary Analysis. Journal of Geographical System, **2(3),** 221-241.

Khalili, N., Wood, J. and Dykes, J**.** (2009) Mapping geography of social networks, Proceedings of the GIS Research UK 17th Annual Conference, (Fairbairn, D., Eds.), pp. 311-315, University of Durham, Durham, UK.

Li, H., Srihari, R., Niu, C. and Li, W. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In Workshop on the Analysis of Geographic References, Edmonton, Canada.

Montello, D.R., Goodchild, M.F., Gottsegen, J., and Fohl, P**.** (2003) Where's Downtown?: Behavioural Methods for Determining Referents of Vague Spatial Queries . Spatial Cognition and Computation, **3(2&3)**, 185-204.

McCurley, S**.** (2001) Geospatial mapping and navigation of the web. In Proceedings of the 10th International WWW Conference Hong Kong, 221-229.

Purves, R. Dykes, J., Edwards, A., Hollenstein, L., Mueller, D., and Wood, J. (2009) Describing the space and place of digital cities through volunteered geographic

information. GeoViz Workshop on Contribution of Geovisualization to the concept of the Digital City, Hamburg, Germany.

Popescu, A., Grefenstette, G. and Moëllic, P. (2008). Gazetiki: Automatic Creation of a Geographical Gazetteer. International Conference on Digital Libraries, 85-93.

Rauch, E., Bukatin, M. and Baker, K. (2003) A confidence-based framework for disambiguating geographic terms. In Workshop on the Analysis of Geographic References, Edmonton, Alberta, Canada.

Silva, M.J., Martins, B., Chaves, M., Cardoso, N. and Afsono, A. (2006) Adding Geographic Scopes to Web Resources. Computers Environment and Urban Systems, 378-399**.**

Smith, D.A. and Crane, G. (2002) Disambiguating Geographic Names in a Historical Digital Library. Proceedings of the 5th European conference on research and advanced technology for digital libraries, London, UK. Springer-Verlag, 127-136.

Smith, D.A. and Mann, G. (2003) Bootstrapping toponym classifiers. In Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL), Workshop on analysis of geographic references, Morristown, NJ, USA, 45-49.

Waters, T. and Evans, A. (2003) Tools for web-based GIS mapping of a “fuzzy” vernacular geography. GIS Research UK (GISRUK), 9-11.

## Biographies

Nazanin Khalili Shavarini is a PhD Candidate at the giCentre, City University London with research interests in visualization and geo social networks. She graduated from Tehran Azad University in 2005 in Computer Hardware Engineering and has an MSc in Information Systems and Technology from City University London.

Dr. Jo Wood is a Reader in Geographic Information at the giCentre at City University London with research interests in geovisualization, terrain modelling and object oriented programming for spatial sciences.

Dr. Jason Dykes is a Senior Lecturer at the giCentre, City University London undertaking applied and theoretical research in, around and between information visualization, interactive analytical cartography and human-centred design.

# The Impact of Target Hardening Policy on Spatial Patterns of Urban Crime in Leeds

Christopher Thompson[1], Mark Birkin[1], Fiona McLaughlin[2], Simon Hodgson[2]

[1] School of Geography, University of Leeds, LS2 9JT
Tel. (0 in UK) 113 34 33300
Christopher Thompson: gy08cpt@leeds.ac.uk / www.geog.leeds.ac.uk/people/CThompson
Mark Birkin: M.H.Birkin@leeds.ac.uk / http://www.geog.leeds.ac.uk/people/m.birkin/

[2]Safer Leeds Partnership, PO Box 612, Leeds West Yorkshire, LS2 7WH

KEYWORDS: Burglary, Target Hardening, Leeds, Displacement, Clustering.

## 1. Introduction

In the wake of the current economic recession, tighter budgets and thinning resources in the public sector have highlighted the need to formulate proactive and prudent ways of allocating effective crime prevention resources within the urban environment. Consequently, in this paper we report the outcomes of collaborative research with government body Safer Leeds. As such, the work adds more evidence to a growing body of literature on spatial patterns in urban crime (Hirschfield et al 1995; Craglia et al, 2000; Rengert and Wasilchick, 2000; Ratcliffe 2004). A distinctive feature of the current work is that we focus on the displacement effect of policy and on temporal variations in criminal activity throughout the year. In particular, we examine the effect of target hardening, the process in which resources are directed to physically increase the security of properties in locations judged to be of notable risk of burglary (e.g. widow bars). For the distribution of crime can be likened to that of iron filings held in a magnetic field. Change the field and the filings are rearranged. As magnets create a force field, so do policy and practice create a crime pattern. As we choose a field by positioning magnets, we "choose" a crime pattern by selecting particular policies and practices (Barr and Pease, 1990). Therefore, an ultimate policy outcome through the partnership with Safer Leeds could be the development and delivery of more effective target hardening strategies within the Leeds district, an area positioned 9th worse out of all 484 local authority districts with 20.5 burglaries per 1000 people (Home Office, 2009).

## 2. Methodology

### *2.1 Data*

Safer Leeds provided individual level burglary data from 2003-2008. The data had information on the date and time the crime was committed, a full address of the property and the geographical coordinates. In addition, Safer Leeds also supplied target hardening figures from 2003-2007 at an individual property level. However, unfortunately this data did not come with the easting and northing values of the properties. In general, both datasets from Safer Leeds were reasonably consistent; nevertheless there was some considerable manipulation required. For example, time of day was removed due to the large degree of uncertainty created by the absence of the owner when many property offenses take place (Ratcliffe, 2004; Ratcliffe, 2006). Furthermore, to ensure a level of consistency and to allow for comparisons, both the crime and target hardening data were aggregated to larger geographies. As the postcode was the common geography, both datasets were amended to show the total number of crimes / houses secured in each year for each postcode. The postcodes in

turn were aggregated to output area level to provide more uniformity in terms of population size, shape and social characteristics. Finally, a number of other datasets were integrated such as the number of houses per output area to help control for the population at risk (Andreson, 2006; Levine, 2006), and corresponding geodemographic codes to provide extra context.

### *2.2 Analytical Techniques*

Once the data had been successfully cleaned and geocoded, a number of different clustering algorithms were utilised to test for burglary hotspots within the data. Firstly, as a way of testing for the significance of clusters in the entire data set, Nearest Neighbour statistic and Ripley's K function global view clustering algorithms were run in ArcGIS 9.2. The choice for using the Nearest Neighbour was based upon its simplicity and efficiency (Clark and Evans, 1954), whereas the K-function is argued to be free from the modifiable areal unit problem (Anselin et al, 2004; Lu and Chen, 2007). In conjunction, various local view algorithms were implemented to identify the actual location of burglary and target hardening hotspots. To guarantee a high level of accuracy a total of three different techniques were used and compared. These were the Getis Ord/Gi* statistic, Geographical Analysis Machine (GAM) developed by Openshaw (1988) and the Kernel Density algorithm.

## 3. Results and Discussion

### 3.1 Spatial

The number of burglaries has increased steadily since 2005, perhaps because of the increase in potential targets for offenders due to a rising population (Leeds City Region, 2009); although current rates are much lower than earlier in the decade (2003/04). The level of target hardening has tended to follow a similar pattern to the number of burglaries. For example, a year with high crime is typically simultaneous with increased target hardening. For exploratory spatial analysis of the data, global clustering algorithms were implemented to test for the existence and significance of clusters. The Nearest Neighbour and Ripley's K function both found significant clustering at a 99% confidence. The location of these clusters was then discovered using the various local view algorithms. Figure 1 displays the results of the Getis Ord Gi*. Those areas with the highest levels of burglary are concentrated towards the central parts of the city, thus pointing towards a typical urban / rural divide (Barr and Pease, 1990; Home Office 2009). Specifically, Headingley, University, Burmantofts, Harehills, and to an extent Richmond Hill, Weetwood, Armley and Kirkstall are those areas most likely to suffer from burglary. High burglary levels in Headingley and University are associated with high student populations which provide easy targets (Craglia et al, 2000), while the other areas represent inner city neighbourhoods with high levels of multiple deprivation (Leeds City Council, 2007).

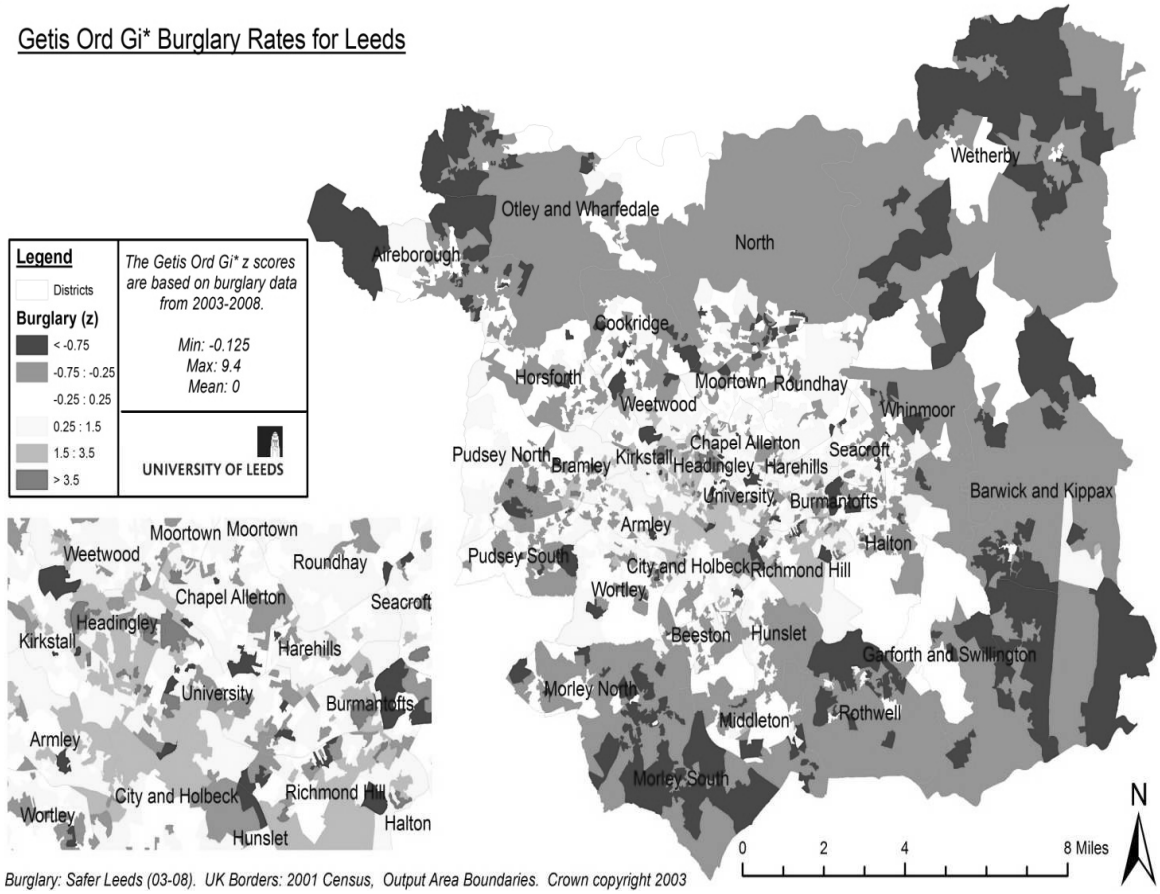



**Figure 1** – Getis Ord Gi* of burglary rates in Leeds

## 4.2 Temporal

Albeit incredibly important, the location of crime only represents a static measure of offending. Therefore, additional analysis was undertaken on the time of year burglaries are committed in Leeds. It was found that burglaries are highest in the winter months (Dec, Jan, and Feb) on account of increased levels of darkness, victims leaving homes vacant over Christmas, and there being new expensive goods in homes to steal. After winter, the rates stay high in spring (Mar, Apr, and May) before reaching a low in summer then rising again in the autumn. Summer often receives the least amount of burglaries because, "…school holidays mean that more people are at home which reduces the opportunity to offend" (Farrell and Pease, 1994, pp.494). Figure 2 helps delineate the spread of crime through the year spatially. Autumn and spring present very similar patterns to Figure 1, however winter and summer have more distinct results. For example, in winter and summer Figure 2 displays particularly high levels of burglary around Headingley as a result of students returning home when term time finishes, leaving houses unoccupied. Moreover, during the summer, the intensity of crime in Burmantofts, Chapel Allerton and Harehills is higher than any other time of the year. Contrary to many, Weisel (2002) believes that in some places, burglary rates can be at their highest in August as residents often leave windows/doors open which increases the opportunity for theft. Furthermore, it could also be a result of families going away on summer holidays, as the areas mentioned contain a large percentage of 'working class' families with young children still in school. This point is supported by the fact burglary patterns revert back again to the normal spatial pattern found in Figure 2 when children return back to school in autumn.

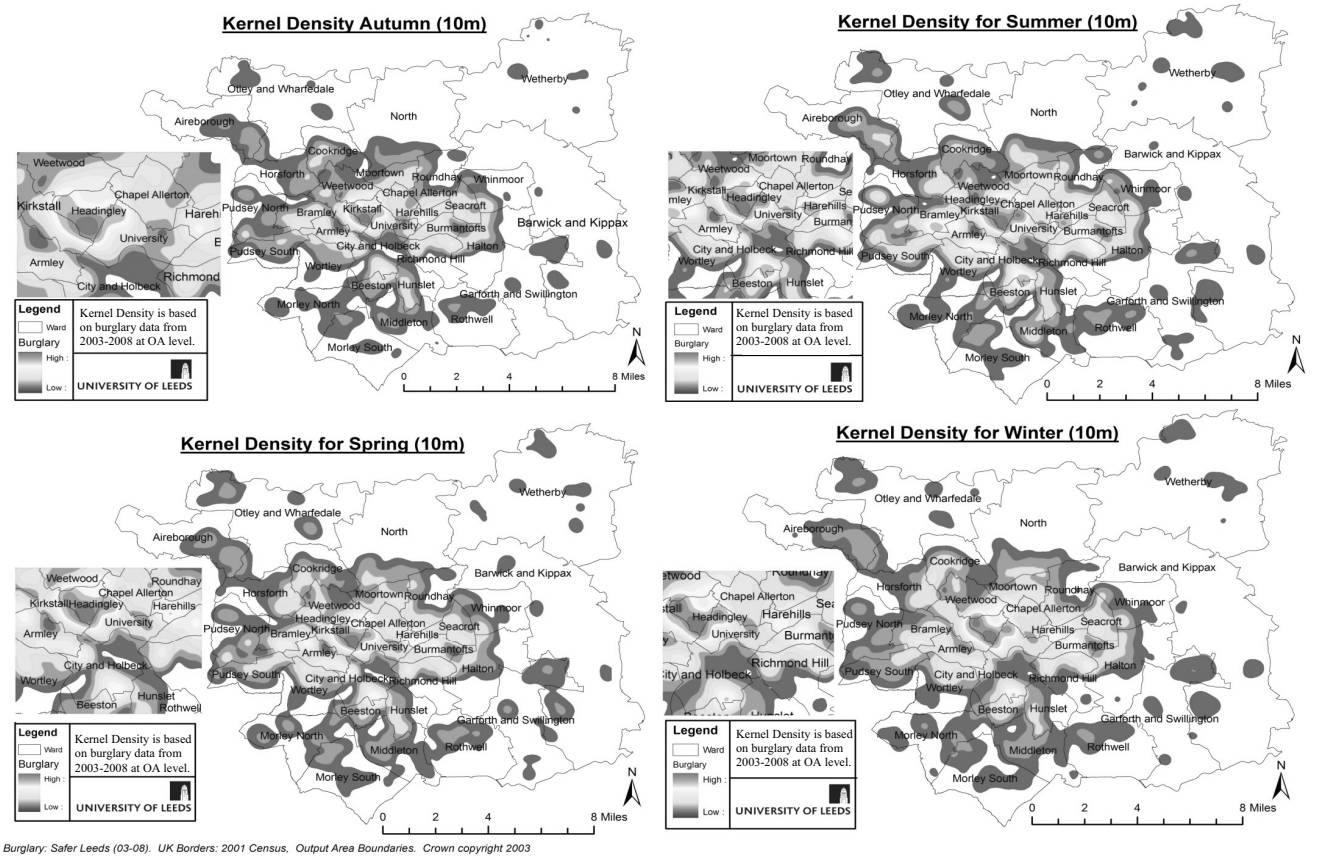



**Figure 2** – Kernel Density Estimation by Season

### 4.4 Displacement

The final part of the analysis and discussion involved searching for any burglary displacement or 'free rider' effects which may have occurred as a result of target hardening in Leeds. For it is argued that crime prevention methods merely deflect crime from the targeted area, to other, more vulnerable places (Meithe, 1991; Town, 2002). Due to the difficulty in finding evidence of crime deflection, only those areas with the worst crime and most intense target hardening were analysed. The year on year analysis showed areas with high levels of target hardening in 2005 and 2007 saw increased crime in neighbouring wards the following year. In 2005, crime shifted slightly from southern parts of Bramley to Armley, from Headingley to Chapel Allerton, from University to Richmond Hill, and from Harehills to Burmantofts. Furthermore, in 2007 heavy target hardening in Southern Roundhay, City and Holbeck, parts of Harehills and Southern Burmantofts appeared to have spread crime into the central parts of Burmantofts, Harehills and Beeston. Despite this, the majority of the years showed no major signs of displacement. In fact, target hardening in 2004 and 2006 caused a spatial reduction in burglaries the following year in 2005 and 2007 respectively. In particular, areas within Headingley, Chapel Allerton, Burmantofts and Seacroft were shown to benefit from 'free rider effects' (reduction in crime within the surrounding area due to target hardening). Subsequently, the implication for broader crime prevention strategy is burglary deflection should be seen as a predictable effect of policies and as a manipulative tool of crime control (Barr and Pease, 1990).

## 5. Future Research

It is advised that future research should be aimed at finding more effective means of measuring burglary displacement, because this still remains an area with many unanswered questions. Furthermore, specific attention should also be paid to the temporal aspects of burglary as a number of studies now exist on urban spatial patterns. It is understood that this would require more accurate data to be recorded and more intra agency collaboration. However, as poor economic conditions act as a catalyst for increased burglary rates and financial investment decreases, there is no better time to identify new ways of reducing crime in a manner which is cost effective, proactive and time efficient.

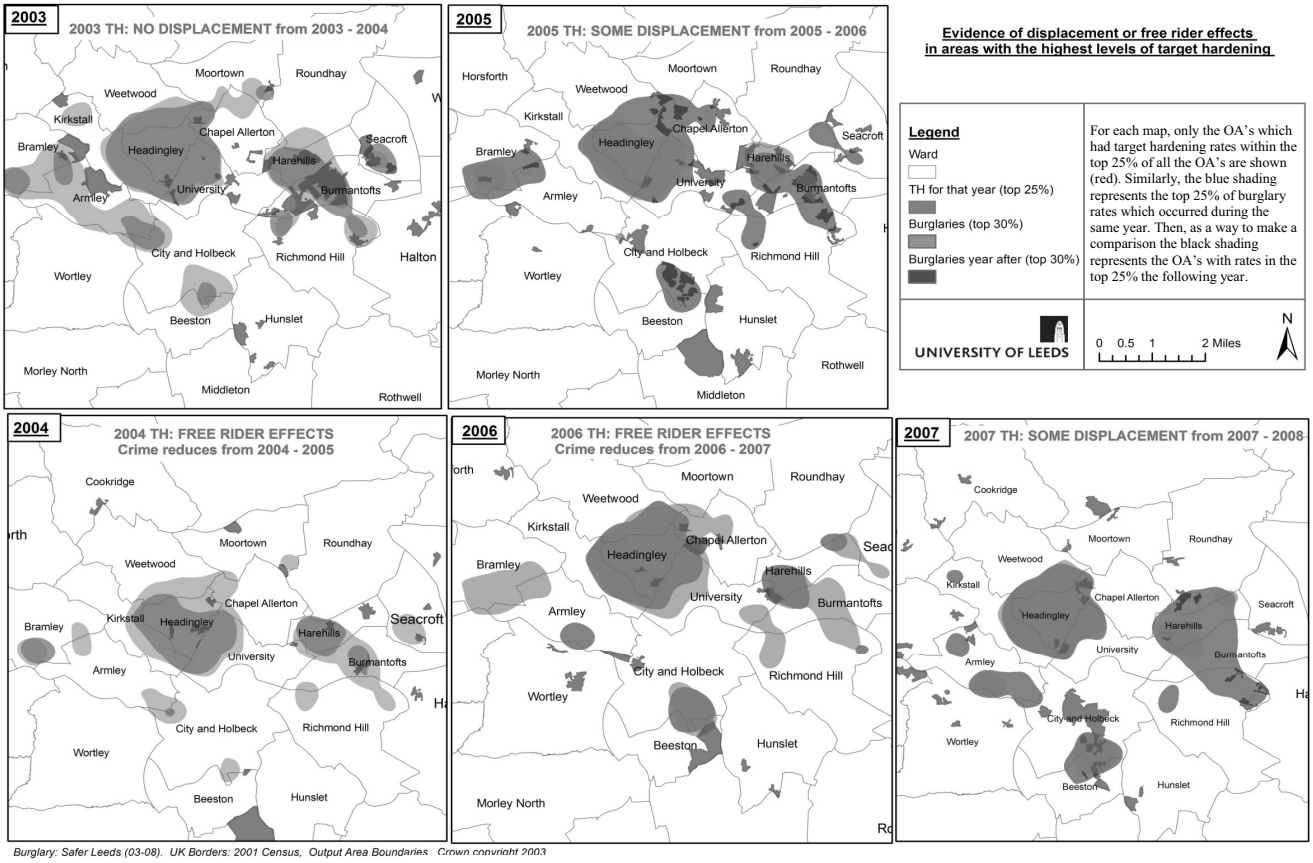



**Figure 3 –** Annual Displacement

**Acknowledgements**

Firstly, I would like to thank Safer Leeds for providing the data to undertake this joint paper. In particular Simon Hodgson and Fiona McLaughlin, who gave up their time to produce the original datasets, discuss aims and answer general queries. Furthermore, many thanks also go to Professor Mark Birkin and Dr Linda See for their continuous support during various stages of the research.

**References**

Andreson, M. (2006) Crime Measures and the spatial analysis of criminal activity, *British Journal of Criminology*, 46, pp.258-285.

Anselin, L., Cohen, J., Cook, D., Gorr, W. and Tita, G. (2004) Spatial Analysis of Crime, *Criminal Justice,* 4, pp.213-262.

Barr, R, and Pease, K. (1990) Crime placement, displacement and deflection, In Tonry, M. and Moms, N. (eds), Crime and Justice: An Annual Review of Research, 12, University of Chicago Press, Chicago.

Clark, P. and Evans, F. (1954) Distance to Nearest Neighbour as a Measure of Spatial Relationships in Populations, *Ecological Society of America,* 35, (4), pp.445-453.

Craglia, M., Haning, R, and Wiles, P. (2000) A comparative evaluation of approaches to urban crime patterns analysis, *Urban Studies*, 37, (4), pp.711-729.

Farrell, G. and Pease, K. (1994) Domestic Disputes and Residential Burglary in Merseyside 1988—98, *British Criminology Journal*, 34, (4), pp.487-498.

Hirschfield, A., Bowers, K. and Brown, P. (1995) Exploring Relations between Crime and Disadvantage on Merseyside, *European Journal on Criminal Policy and Research*, 3 (3), pp.93-112.

Home Office (2009) Atlas

http://www.homeoffice.gov.uk/rds/ia/atlas.html [Accessed online: 26/07/2009]

Leeds City Council (2007) An understanding of the Indices of Deprivation.

(Available online: http://www.leeds.gov.uk/files/Internet2007/2008/week5/inter__a5ccd1c3-8752-4bb3-ad7f-471f10778e7c_ded31a85-f116-4fba-88c8-8f77c57c46cc.pdf)

Leeds City Region (2009) New Growth Point Status for the Leeds City Region *http://www.leedscityregion.gov.uk/uploadedFiles/Research_and_Publications/Housing/H01%20-%20LCR%20FINAL%20New%20Growth%20Points%20bid%20(31Oct07).pdf [Accessed Online: 21/08/2009]*

Levine, N. (2006) Crime Mapping and the Crimestat Program, Geographical Analysis, 1, (1), pp.1-11.

Lu, Y. and Xuwei, C. (2007) On the false alarm of planar K-function when analyzing urban crime distributed along streets, *Social Science Research*, 36, (2), pp.611–632.

Miethe, T. (1991) Citizen based crime control activity and victimization risks: Examination of displacement and free rider effects, *Criminology*, 29, (3), pp.419-439.

Openshaw, S. and Charlton, M. (1988) Searching for Leukaemia Clusters Using a Geographical Analysis Machine, *Regional Science Association*, 65, pp.95-106.

Ratcliffe, J. (2004) The Hotspot Matrix: A Framework for the Spatio-Temporal Targeting of Crime Reduction, *Police Practice and Research, 5, (1), pp.5-23.*

Ratcliffe, J. (2006) A Temporal Constraint Theory to Explain Opportunity-Based Spatial Offending Patterns, *Journal of Research in Crime and Delinquency*, (43), 3, pp.261-291.

Rengert, G. and Wasilchick, J. (2000) Suburban Burglary: A Tale of Two Suburbs, Charles Thomas, Springfield.

Town, S. 2002, Crime displacement: The perception, problems, evidence and supporting theory, Bradford District Council**.** (Available online: www.crimereduction.gov.uk/ skills10.htm).

Weisel, D. (2002) Burglary of Single-Family *Houses, Centre for problem orientated policing*, 18, pp1-4. (Available online: http://www.popcenter.org/problems/burglary_home/).

**Biography**

*Chris Thompson – I have recently started a PhD within the Department of Geography at the University of Leeds. My research is tilted, "Retail Spending and Store Location during a Recession: An Analysis of Changing Consumer Behaviour and Interaction Patterns". Other research interests include spatial modelling within crime, health and population geography.*

*Professor Mark Birkin – I am Professor of Spatial Analysis and Policy in the School of Geography, University of Leeds. My interests include spatial microsimulation, geodemographics, mathematical modelling of urban service provision, as well as GIS, particularly in the context of Spatial Decision Support Systems. I am a founding co-editor of the Journal Applied Spatial Analysis and Policy.*

*Fiona McLaughlin - Principal Partnership Analyst within the Strategic Intelligence Unit at Safer Leeds.*

*Simon Hodgson - Performance Review Manager within the Safer Leeds Partnership Team*

# Street-level Spatial Scan Statistic for Crime Hotspot Detection

Shino Shiode[1], Narushige Shiode[2]

[1]Department of Geography, Environment and Development Studies – Birkbeck, University of London
Malet Street, London WC1E 7HX, UK
Email: s.shiode@bbk.ac.uk

[2]School of City and Regional Planning – Cardiff University
Glamorgan Building, King George VII Avenue, Cardiff CF10 3WA, UK
Email: ShiodeN@cardiff.ac.uk

KEYWORDS: Crime Hotspot Detection, Network, Spatial Scan Statistic, Street Crime

## 1. Introduction

In the field of crime science, crime hotspot detection techniques have been significantly refined over the past few decades in response to the strong demand from the academic community as well as the professionals at law enforcement agencies for a more precise identification of where spatial concentrations of crime occur.

One of the recent trends in the studies of the spatial distribution of crime incidents in general, and crime hotspots in particular, is the change in the scale of analysis from the aggregate level to the disaggregate level. In particular, studies of micro crime places have been rigorously pursued in criminology thus drawing the attention of analysts and professionals to the analyses using a more micro-scale geographic unit such as street blocks and street segments (Weisburd *et al*. 2009). However, the current range of crime hotspot detection methods is not necessarily suitable for the detailed, micro-scale analysis of crime distributions. This presents a problem, as one of the main interests in crime science is to identify the concentration of crime derived through the interpretation of the exact location of the crime incidents, whose spatial distribution and the hotspot locations are often regulated by the configuration of the street network of the urban environment. Although many studies point this out (e.g. Weisburd and Eck 2004, van Wilsem 2009, Weisburd *et al*. 2009), the street network has hardly been incorporated into the actual analysis of crime hotspots, except for few studies such as Okabe *et al*. (2009).

This study aims to improve the quantitative description of the spatial patterns of street-level crime hotspots by incorporating the network distance into the analysis. The study first presents a new type of hotspot detection method which adopts and extends, to the network space, one of the most frequently used search-window-type methods. The study then conducts empirical analysis with simulated data for testing the validity of the proposed method as well as the robustness of the program that executes the proposed method. The study concludes with a discussion of main findings and the validity of the use of the proposed method in detecting the street-level hotspots.

## 2. Network-Based Search Window and Network-Based Spatial Scan Statistic (NT-SaTScan)

Conventional search-window methods are usually carried out by exhaustively sweeping across the study area with a search window (a circular or an ellipsoidal shaped sub-area) in such a way that the window constantly covers a fixed amount of area while enclosing different sets of point observations. This study applies the concept of the search window to network space where the sub-area search window is replaced by a window that consists of a sub-network, or a collection of line segments. As in the case with the circular search window, it moves along the network across the entire study area and detects point observations within the extent of the sub-network search window. Using this network-based search window (NT-SW), this study introduces a network-based search-window-type hotspot detection method by developing the network equivalent of *Spatial Scan Statistic* (Kulldorff 1997).

Since the homogeneous Poisson process is assumed for *the network-based Scan Statistic* (NT-SaTScan) in this study, the analysis is bound by the total number of crime incidents *n.* Extending the process of hotspot detection of the conventional Spatial Scan Statistic (PL-SaTScan) to the analysis of network space, the spatial scan statistic $S_{NT}$ defined on the network is the maximum likelihood ratio over all possible $N_{sw}$, i.e.

$$S_{NT} = \max_{N_{sw}} \left\{ \frac{L(N_{sw})}{L_0} \right\}, \tag{1}$$

where $L(N_{sw})$ is the maximum likelihood for the network search window $N_{sw}$ and $L_0$ is the likelihood function under the null model. Let $n(N_{sw})$ be the number of incidents observed in $N_{sw}$ and $\mu(N_{sw})$ be the expected number of incidents within $N_{sw}$ under the null model, so that $\mu(N_{sw})=\lambda$ and $\mu(N)=n$. Then, the likelihood ratio for a specific search window is

$$\frac{L(N_{sw})}{L_0} = \left\{ \frac{n(N_{sw})}{\mu(N_{sw})} \right\}^{n(N_{sw})} \left\{ \frac{n - n(N_{sw})}{n - \mu(N_{sw})} \right\}^{n-n(N_{sw})}, \text{ if } n(N_{sw}) > \mu(N_{sw}), \text{ and}$$

$$\frac{L(N_{sw})}{L_0} = 1, \text{ otherwise.} \tag{2}$$

By maximising the likelihood ratio over all $N_{sw}$, the single $N_{sw}$ that constitutes the most likely cluster can be identified.

## 3. Analysis

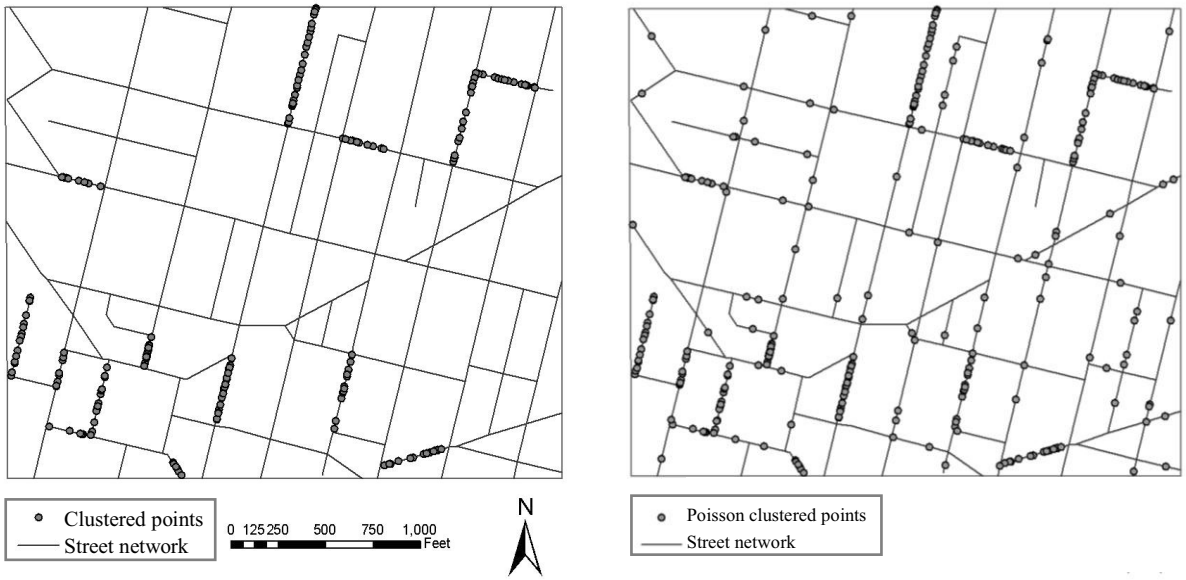
### 3.1 Simulated Clustered Point Pattern

A series of simulated distributions which follows *the Poisson cluster process* (Diggle 2003) is generated along the network for the purpose of creating an artificial point distribution whose hotspot locations are known, and then using it for comparing the relative performance of the proposed method to that of their conventional counterparts in extracting these hotspots. Each distribution is produced by (1) generating 200 random points on 14 line segments randomly selected from the sample street network (Figure 1(a) shows an illustrative example); (2) generating 100 random points on the entire network; (3) joining them to form an inhomogeneous clustered point pattern consisting of 300 points (Figure 1(b)).

### 3.2 Application of the PL-SaTScan and the NT-SaTScan

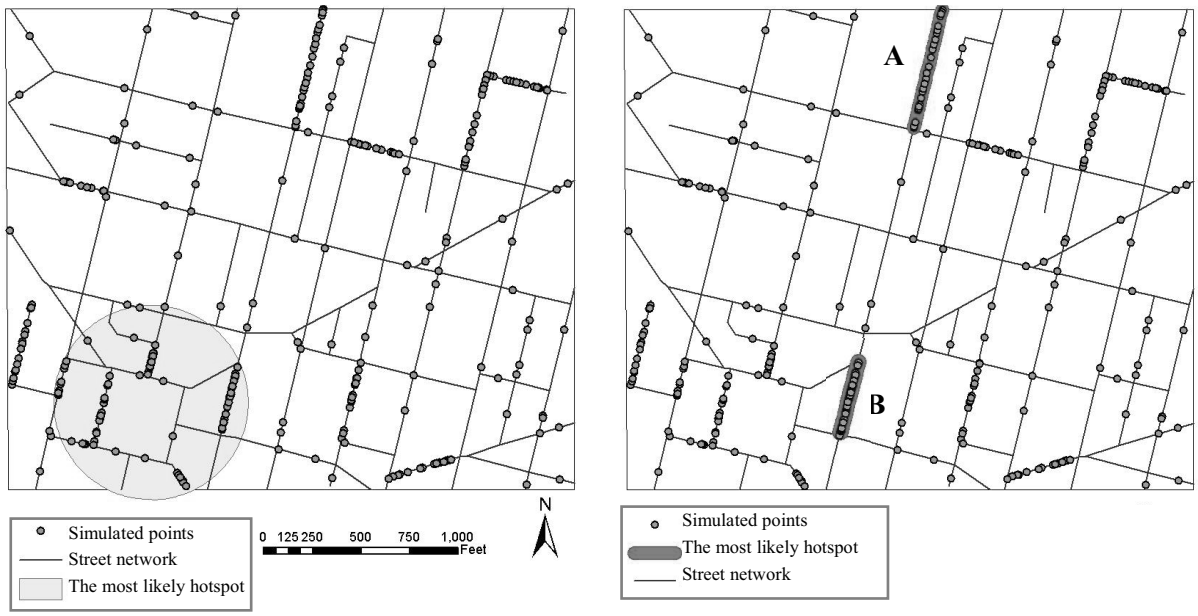
In order to conduct empirical analysis using the NT-SaTScan, the method was coded as proprietary computer programs. The objective is to detect the most likely area of crime concentration, and to this end, both the PL-SaTScan and the NT-SaTScan are applied to the point distribution of the simulated Poisson cluster point process. The circular area detected by the PL-SaTScan covers a relatively wide area including multiple cluster locations (Figure 2(a)). In contrast, the NT-SaTScan detects only a single line segment that corresponds to one of the 14 line segments (location A in Figure 2(b)) on which the clusters were created by simulation. This illustrates that the NT-SaTScan can identify the concentration of crime incidents in a more concise manner than the PL-SaTScan could.

In order to examine the relative performance of the two SaTScans, the simulated cluster point distribution is modified with the introduction of 20 new points in addition to the original Poisson cluster distribution. All of the new points are added to one of the 14 cluster locations (i.e. location B in Figure 2(b)) to intentionally create an area with the highest point density. With this modified point distribution, the PL-SaTScan produces exactly the same result as before, whereas the NT-SaTScan detects the line segment that has received the additional 20 points.




(a) (b)

**Figure 1.** Construction of a simulated distribution: (a) random generation of 200 points on the 14 line segments, and (b) Poisson cluster distribution of 300 points.




(a) (b)

**Figure 2.** Detected most likely hotpots by applying (a) PL-SaTScan and (b) NT-SaTScan, where the hotspot on the network was detected at location A in the original distribution and at location B in the modified distribution.

## 4. Conclusion

This study presented the network-based spatial scan statistic that can be applied to the analysis of the disaggregate distributions of crime incidents observed on a network. The area detected by the PL-SaTScan is indeed the most likely hotspot on the plane as there is a concentration of 5 clustered locations which together forms a large cluster area. However, when it comes to detecting the exact location of the hotspot along the street network, the NT-SaTScan is expected to be more precise and is also sensitive to the local changes in the cluster locations, and this was confirmed by successfully detecting the new hotspot that achieved the highest point density.

The findings from this study only offer an indicative, rather than conclusive, evidence regarding the relative advantage of the NT-SaTScan over the PL-SaTScan. However, given the results of the analyses on the simulated distribution, it would not be unreasonable to assume that the network-based methods can offer a more accurate and sensitive outcome in detecting hotspots, especially when we look at the finer and more local street-level distributions.

## 5. Acknowledgements

I greatly appreciate Professor Robert Haining for his valuable comments in this study. This study is supported by the Canon Foundation in Europe and the International House of Japan.

## References

Diggle, P.J. (2003) *Statistical Analysis of Spatial Point Patterns*. New York: Oxford University Press.

Kulldorff, M. (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods* **26**, pp.1481-1496.

Okabe, A., Satoh, T. and Sugihara, K. (2009) A kernel density estimation method for networks, its computational method and a GIS-based tool, *International Journal of Geographical Information Science*, **23**(1), pp.7-32.

van Wilsem, J. (2009) Urban streets as micro contexts to commit violence, In Weisburd, D., Bernasco, W. and Bruinsma, G.J.N. (eds.), *Putting Crime in Its Place: Units of Analysis in Geographic Criminology*. Springer Verlag, New York.

Weisburd, D., Bernasco, W. and Bruinsma, G.J.N. (2009) *Putting Crime in its Place: Units of Analysis in Geographic Criminology*. Springer Verlag, New York.

Weisburd, D. and Eck, J. (2004) What can police do to reduce crime, disorder and fear?, *The Annals of the American Academy of Political and Social Science* **593**, pp.42-65.

## Biography

Shino Shiode is a lecturer in GIScience at Department of Geography, Environment and Development Studies, Birkbeck College, University of London. Her research interests are methodology of spatial analysis with a particular focuses on crime, health and urban applications.

Narushige Shiode is a lecturer in Spatial Analysis and GIS at School of City and Regional Planning, Cardiff University. His research interests are methods of spatial analysis and modelling, urban growth dynamics, and cartography and geo-visualisation.

# Exploring intra-urban variations in the incidence of fire events using a geodemographic classification

Tessa Anderson[1], Jonathan Corcoran[2], Gary Higgs[3]
[1]The Kadoorie Institute, The University of Hong Kong, Pokfulam Road, Hong Kong
Tel. +86 15889500401 | Email. Tessa81@hku.hk
[2]Department of Geography, Planning and Environment Management, University of Queensland, Brisbane, Australia

[3] Faculty of Advanced Technology, University of Glamorgan, Pontypridd, UK

KEYWORDS: fire, geodemographics, community, public policy and GIS

## 1. Introduction

Aims

- to advance our previous research which has looked at potential associations between fire incidence and standard deprivation indices and unconstrained census variables to see if geodemographic classifications have potential in explanatory as well as predictive applications
- to identify neighbourhoods potentially at risk from different types of fire events based on their geodemographic classification in Cardiff, South Wales.
- to demonstrate their potential in education and safety campaigns through an examination of spatial variations in neighbourhood patterns

## 2. Use of Geodemographic classifications in public sector applications

Geodemographic classifications are created by performing clustering analysis of a combination of census of population, survey and commercially motivated lifestyle information in order to permit a discrimination of neighbourhood areas at detailed spatial scales based on unit postcode or census output area building blocks (Harris et al., 2005). Traditionally these measures have been used in the retail sector for a number of commercially driven aims (Longley and Clarke, 1995). However, with the active marketing of such classifications to public sector agencies, there is a growing literature on the potential use of geodemographic classifications outside the traditional use of these methods in the private/commercial sector (Brown et al., 2000 Aveyard et al., 2002; Farr et al., 2008; Tickle et al., 2000 Batey and Brown, 2007; Batey et al. 2008; Petersen et al., 2009; Longley et al., 2006; Longley and Singleton, 2009).

Application in fire and emergency services: a number of fire and rescue services in the UK are using geodemographic classifications in combination with historical records of fire incidence within fire safety teams to identify relationships between fire incident types and neighbourhood groups and help inform community safety initiatives (Smith et al., 2008). This has involved for example targeting areas for fire reduction initiatives through a classification for example of fire dwelling rates for groups based on the classification. For example, with regard to the Experian MOSAIC™ classification, research has shown that fire incidence is higher for people who live in areas described as,

- People living in social housing with uncertain employment in deprived areas
- Low income families living in estate based social housing
- Older people living in social housing with high care needs

Geodemographic analysis is increasingly being used alongside other techniques with supporting information including local knowledge of fire incidence. The study by Smith et al (2008) used

dwelling fire data at postcode level. They used a geodemographic classification to allocate risk categories to groups based on the individual factors found to be associated with increased risk. Thus a qualitative approach was used to provide a subjective fire risk ranking to each geodemographic group.

## 3. Methodology

Mosaic UK classifies all UK consumers into 61 neighbourhood types arranged into a hierarchy of 11 groups which describe their socio-economic and socio-cultural behaviour based on the 1.4 million unit postcodes. Just over 400 variables were used to create the current version of Mosaic based on the publication of the 2001 UK Census. In the first instance an index score was established for each postcode in Cardiff. Each fire event (and its postcode) were allocated a Mosaic Type (and Group), and from this and the household count for Cardiff it was possible to establish a 'risk' surface of index scores for Cardiff (an index score of 100 indicates the 'expected' level of risk, whereby any score of this, has a higher than average propensity).. Figure 1 highlights the spatial nature of Mosaic Groups in the centre of Cardiff

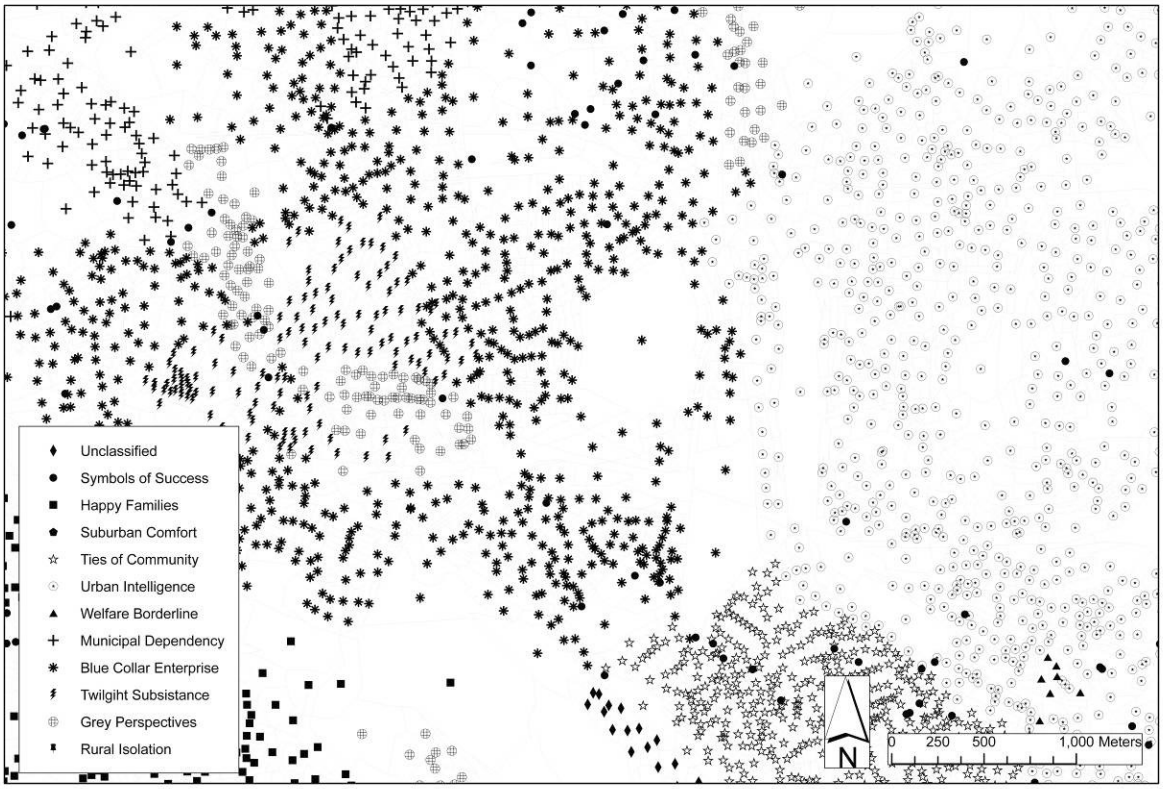



**Figure 1:** Map of the distribution of MOSAIC Groups in Cardiff

## 4. Results

| Mosaic Type | Cardiff Households | Cardiff % households | Hoax | Secondary | Vehicles | Buildings | All fires |
|---|---|---|---|---|---|---|---|
| E31 Caring Professionals | 20027 | 7.89 | 0 | 2 | 1 | 1 | 94 |
| C19 Original Suburbs | 17777 | 7.00 | 62 | 123 | 95 | 106 | 101 |
| H46 White Van Culture | 17073 | 6.72 | 3 | 4 | 5 | 2 | 94 |
| A05 Provincial Privilege | 15277 | 6.02 | 22 | 63 | 40 | 52 | 97 |
| E33 Town Gown Transition | 14406 | 5.67 | 11 | 77 | 72 | 15 | 98 |
| H47 New Town Materialism | 11365 | 4.48 | 122 | 26 | 35 | 32 | 96 |
| D21 Respectable Rows | 10246 | 4.04 | 22 | 3 | 4 | 13 | 95 |
| C18 Sprawling Subtopia | 8024 | 3.16 | 83 | 144 | 116 | 115 | 102 |
| E34 University Challenge | 7712 | 3.04 | 13 | 42 | 29 | 4 | 96 |
| G41 Families on Benefits | 7459 | 2.94 | 101 | 17 | 24 | 25 | 96 |
| F37 Upper Floor Families | 6542 | 2.58 | 0 | 0 | 0 | 0 | 94 |
| D24 Coronation Street | 6506 | 2.56 | 3 | 19 | 30 | 4 | 95 |
| D27 Settled Minorities | 6419 | 2.53 | 8 | 6 | 2 | 3 | 94 |
| A06 High Technologists | 6398 | 2.52 | 96 | 82 | 146 | 184 | 100 |
| G42 Low Horizons | 6089 | 2.40 | 79 | 64 | 95 | 69 | 98 |
| E29 City Adventurers | 5909 | 2.33 | 29 | 60 | 70 | 54 | 97 |
| B11 Families Making Good | 5823 | 2.29 | 288 | 58 | 130 | 227 | 101 |
| A03 Corporate Chieftains | 5590 | 2.20 | 64 | 103 | 93 | 87 | 100 |
| B09 Fledgling Nurseries | 5376 | 2.12 | 220 | 70 | 143 | 227 | 101 |
| B13 Burdened Optimists | 4579 | 1.80 | 41 | 113 | 114 | 78 | 100 |
| B12 Middle Rung Families | 4382 | 1.73 | 0 | 2 | 6 | 2 | 94 |
| J52 Childfree Serenity | 3874 | 1.53 | 0 | 0 | 0 | 0 | 94 |
| A04 Golden Empty Nesters | 3766 | 1.48 | 82 | 373 | 226 | 85 | 111 |
| B10 Upscale New Owners | 3501 | 1.38 | 64 | 27 | 39 | 36 | 96 |
| C16 Conservative Values | 3402 | 1.34 | 196 | 276 | 174 | 188 | 108 |

**Table 1:** Index Scores for different fire types in Cardiff

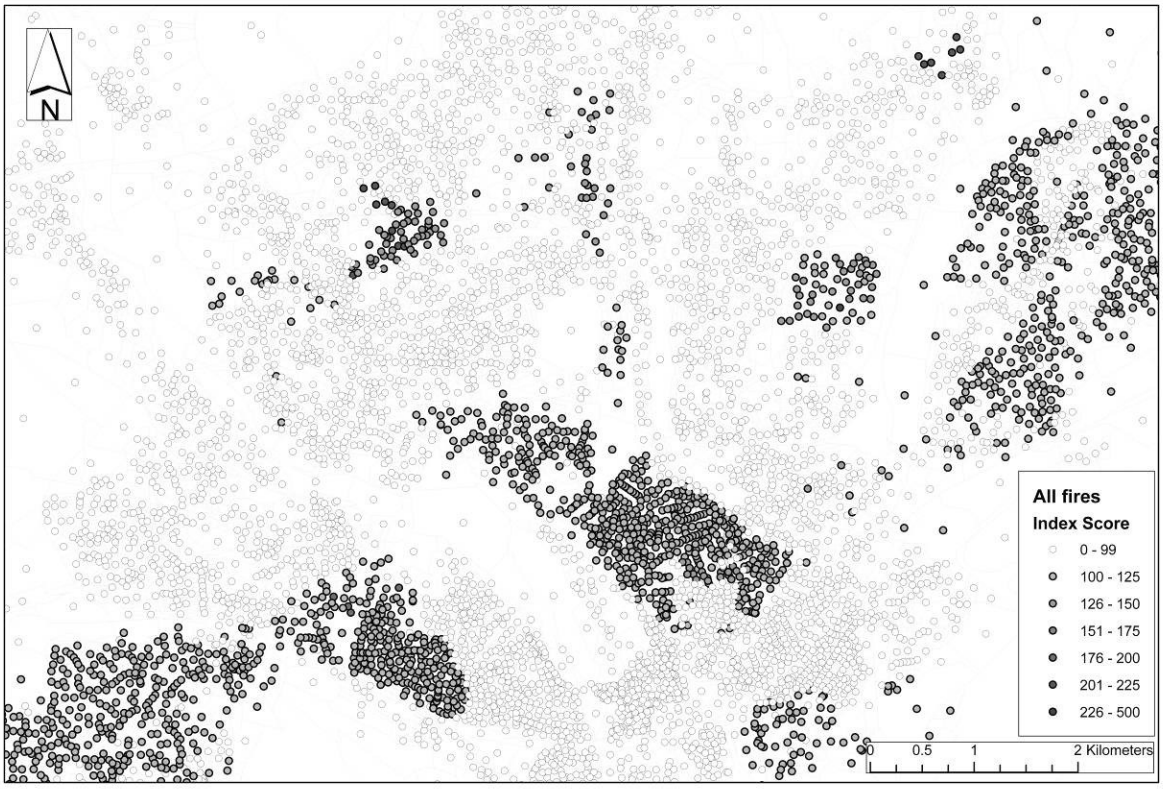



**Figure 2:** Map of the index scores for all fires in Cardiff Centre

The results indicate some patterns of fire risk within Cardiff. Table 1 suggests some interesting variation, for example Mosaic Type H4 New Town Materialism has a greater likelihood of suffering

hoax fire events that any other type of fire event. Figure 2 shows the patterning of fire risk for all fires in Cardiff.

## 5. Discussion

There are a considerable number of fire services in the UK which use geodemographics to profile their areas (notably, Cambridgeshire and Cleveland). One of the key concerns amongst operational users is the ability to use geodemographics not only to profile the areas most at risk but to use it, to effectively engage with the residents of these areas to ensure reduction of fires. Geodemographics can be used to identify areas of priority engagement and the types of engagement most appropriate. The results from this paper will be used eventually alongside a user requirements survey of fire services in the UK, understanding how they use geodemographics, for what purpose and outcome. Currently not all fire and rescue services use this tool, and by combining qualitative data it will be possible to understand the policy implications of this tool and how it can be further utilised to reduce fire.

## 6. Conclusion

Geodemographic classifications have the potential, when used in conjunction with some of the other techniques we have been using to investigate trends in fire incidence, to examine spatial variations in incidence by neighbourhood type and be used in for example education campaigns by targeting schools and households within such areas. Furthermore the use of such neighbourhood classifications may provide a route through which factors such as community cohesion and social capital are incorporated into such analysis to provide a fuller picture of the types of factors associated with spatial and temporal variations in fire incidence.

**References**

Ashby DI and Longley PA (2005) Geocomputation, geodemographics and resource allocation for local policing *Transactions in GIS* (9) pp53-72

Aveyard P, Manaseki S and Chamber J (2002) The relationship between mean birth weight and poverty using the Townsend deprivation score and the Super Profile classification system *Public Health* 116(6) pp308-314

Batey P and Brown P (2007) The spatial targeting of urban policy initiatives: a geodemographic assessment tool *Environment and Planning A* 39 pp2774-2793

Batey P, Brown P and Pemberton S (2008) Methods for the spatial targeting of urban policy in the UK: a comparative analysis *Journal of Applied Spatial Analysis* 1 pp117-132

Brown PJB, Hirschfield AFG and Batey PWJ (2000) Adding value to census data: public sector applications of the Super Profiles geodemographic typology *Journal of Cities and Regions* 10 pp19-32

Farr M, Wardlaw J and Jones C (2008) Tackling health inequalities using geodemographics: A social marketing approach *International Journal of Market Research* 50(4) pp449-46

Hamnett C, Ramsden M and Butler T (2007) Social background, ethnicity, school composition and educational attainment in East London *Urban Studies* 44(7) pp1255-1280

Harris R, Sleight P and Webber R (2005). *Geodemographics, GIS and Neighbourhood Targeting.* Wiley, Chichester

Longley PA (2007) Some challenges to geodemographic analysis and their wider implications for the practice of GIScience *Computers, Environment and Urban Systems* 31 pp617-622

Longley PA, Ashby D Webber R and Li C (2006) Geodemographic classifications, the digital divide and understanding customer take-up of new technologies *BT Technology Journal* 24(3) pp76-74

Longley PA and Clarke GP (1995). *GIS for business and service planning*. Wiley, Chichester.

Longley PA and Goodchild MF (2008) The use of geodemographics to improve public service delivery in Hartley J, Donaldson C, Skelcher C. and Wallace M (Eds.) *Managing to Improve Public Services*, Cambridge, Cambridge University Press, pp176-194

Longley PA.and Singleton AD (2009) Linking social deprivation and digital exclusion in England *Urban Studies* 46(7) pp1275-1298

Mayhew P, Maung NA and Mirlees-Black C (1993) *The British Crime Survey* Home Office Research Studies, London

Petersen J, Atkinson P, Petrie S, Gibin, M, Ashby D and Longley P (2009) Teenage pregnancy – new tools to support local health campaigns *Health and Place* 15(1) pp300-307.

Smith R, Wright M and Solanki A (2008) Analysis of fire and rescue service performance and outcomes with reference to population socio-demographics Fire Research Series 9/2008, Department for Communities and Local Government, London.

Tickle M, Brown P, Blinkhorn A and Jenner T (2000) Comparing the ability of different area measures of socio-economic status to segment a population according to caries prevalence *Community Dental Health* 17(3) pp138-144

# Exploring and Mapping Patterns of Vandalism amongst Young People: When is a problem not a problem?

Ellie Bates[1], William Mackaness[2]

[1]Edinburgh Law School, University of Edinburgh, Old College, South Bridge Edinburgh EH8 9YL Tel: 07745 406959 | Email: e.j.w.bates@sms.ed.ac.uk

[2]Institute of Geography, School of GeoSciences, University of Edinburgh, Drummond Street, Edinburgh, EH8 9XP Tel: 0131 650 8163 | Email: william.mackaness@ed.ac.uk

KEYWORDS: vandalism, youth crime, public crime maps, ESDA, geovisualisation

## 1. Introduction

Can there be circumstances when the way in which we classify, standardise and visualise crime data might give us very different results yet both might be correct? This research suggests there can be, and has important practical implications as crime data is increasingly being made available to the public, both as maps and raw data. Whilst seeking to unravel self-reported patterns of vandalism amongst young people, and more generally trying to understand the place and time dynamics of vandalism in a Scottish context, our research has uncovered some interesting issues.

This paper demonstrates the value of the recursive Exploratory Spatial Data Analysis (ESDA) and GIS approaches in analysis and visualisation, but also illustrates that care is needed in the interpretative process. Understanding crime data is not easy; data are often incomplete (with unreported crime sometimes referred to as the dark number), classifications of crime vary between data collection methods and, if the data are to be standardised, careful consideration needs to given to the denominator used. This short study suggests careful best practice guidance on the presentation and interpretation of mapped crime data is required including consideration of the target audience.

## 2. Vandalism is a crime related to place

Crime has been considered to have a significant spatial element by social scientists for at least 150 years; from the nineteenth century France, to work of the Chicago School in the 1920s and 30s, with debates around the broken windows theory from the early 1980s, to the growing interest in GIS and crime mapping and concepts such as crime pattern theory, collective efficacy and situational action theory since the mid 1990s (Melossi 2008, Bottoms, 2007, Wortley 2008, Chainey & Ratcliffe, 2006).

Vandalism for the purposes of this paper is defined as: an act damaging or defacing property; which may be criminal if the property belongs to someone else and the act was intentional. Vandalism is often done to property in a fixed location such as walls, windows or street furniture; it can be seen as a quintessentially spatial crime and inherently tied up with theoretical understandings crime and place. The routine activity approach as originally developed by Lawrence Cohen and Marcus Felson provides a useful framework with which to explore why crimes occur at a particular time and place (Cohen and Felson, 1979). As such GIS and ESDA can play a key role in exploring the place and time elements of this approach.

## 3. A case study

One of the key elements of the routine activity approach is the presence of the motivated or likely offender. In the case of vandalism various theories and typologies have been proposed as to what

might underlie the motivation to commit vandalism (Levy-Leboyer (1984), Goldstein (1996)) and there is a growing literature considering motivations involved in graffiti writing (Halsey & Young 2006). However, there appears to be little research considering whether the likely offender is the same person for different sub-types or styles of vandalism, along with how this might relate to other crimes, and place and time. This research seeks to start to fill this gap using data from the Edinburgh Study of Youth Transitions and Crime (ESYTC).

Vandalism is often thought to be a fairly common activity by younger people. The ESYTC is a self-report longitudinal cohort study whose data is freely available on application. The study followed a group of around 4,300 young people representing around 92% of all children in a year at secondary school across all parts of Edinburgh; as such it has greater breadth that some other similar studies of this type that limit themselves to sections of a city or only males (McVie, 2001, McVie 2003, Smith et al 2001, Smith, 2007). Young people were surveyed in 6 annual sweeps from 1998 when they were aged around 12 until aged around 17. The survey asked them to self-report if they had been involved in behaviour that was criminal or potentially anti-social, and also asked a wide range of other questions about the young person, their school life, home life and neighbourhood. Particularly useful for this project, in 5 out of 6 survey sweeps, separate questions were asked relating to destructive vandalism, graffiti writing and fire-raising (arson).

## 4. Methodology

Research was conducted using a exploratory data analysis and exploratory spatial data analysis approach. As such the work has been recursive and made much use of descriptive graphs and maps and some descriptive statistics. The first 4 sweeps of the study were considered as this was data where postcodes were already attached. Postcodes were then matched to statistical geographies. Priority was given to analysis of sweep 3 as this appeared to be a peak age for vandalism offences (at around age 14) and this sweep also included a number of other useful neighbourhood perception questions. 75.4% of records were successfully geocoded for Sweep 3. These were then aggregated to Scottish Data Zone and Intermediate Geography spatial units and explored using SPSS and the exploratory spatial data analysis package OpenGeoda. Data was mapped using the excess risk mapping tool in OpenGeoda which produces Standardised Offence Ratios (SORs) (Ceccatto & Haining, 2005). These maps visualise the data in comparison to the average rate for the region (in this case the City of Edinburgh) with 4 representing 4 times the regional rate, and 0.25 representing a quarter the regional rate.

## 5. Classification - Graffiti and Vandalism: not the same thing?

Graffiti writing is much more common amongst young people than destructive vandalism or arson. Levels of graffiti writing are highest in sweep 3, with 48.8% young women and 37.0% of young men reporting they have written or sprayed graffiti one or more times (Figure 1).

In Scotland destructive vandalism and graffiti are classified together as ‘vandalism’, in publicly reported police recorded crime, and generally grouped together with housebreaking (burglary) as property crime. This broad classification hides some interesting differences. For example, amongst ESYTC respondents graffiti writing differs from vandalism, fire-raising, housebreaking and violence in that in sweeps 2, 3 and 4 more young women report involvement than men. Analysis found that the strength of the relationship between young people doing both vandalism and graffiti, although statistically significant, was not strong , whilst relationships between vandalism and housebreaking and graffiti and housebreaking were weak (Table 1).

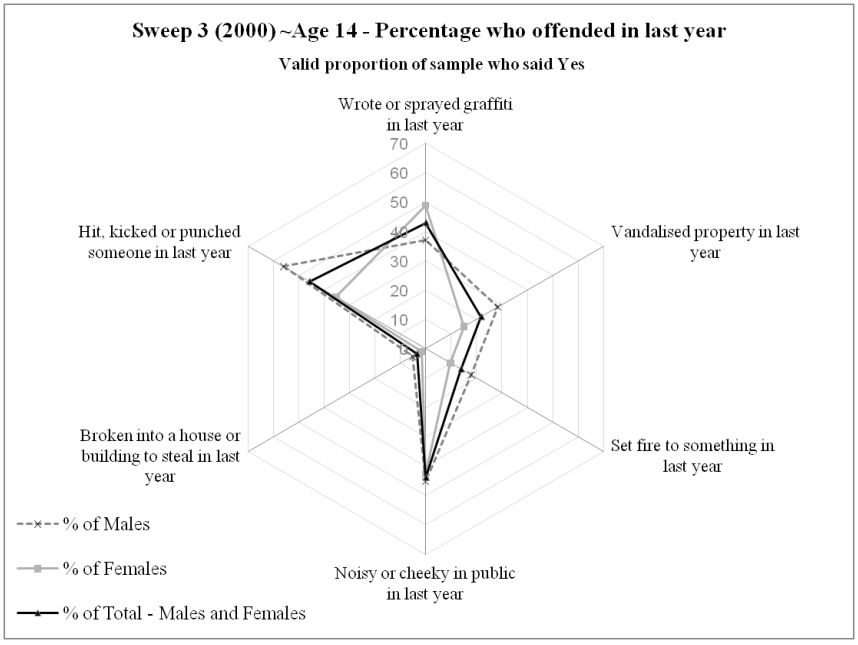



**Figure 1.** Young people reporting involvement in specific crimes ESYTC Sweep 3

**Table 1.** Results of tests to assess if a young person does one crime, do they also do another? (ESYTC Sweep3)

| **Behaviour** | **Pearson $X^2$** | **Cramers V** | **Strength of relationship** |
|---|---|---|---|
| Vandalism and Graffiti | 627.933 | 0.383 | medium |
| Vandalism and Fire-raising | 967.890 | 0.477 | medium/strong |
| Vandalism and Housebreaking | 310.394 | 0.270 | weak/medium |
| Vandalism and Rowdy behaviour | 703.106 | 0.406 | medium |
| Graffiti and Fire-raising | 417.976 | 0.313 | medium |
| Graffiti and Housebreaking | 113.860 | 0.163 | weak |
| Graffiti and Rowdy behaviour | 838.778 | 0.443 | medium |
| all values significant $p<0.001$, Cramers V reports a value between -1 and +1 | | | |

## 6. Visualisation and Standardisation - Truth is a many sided figure

Following from this, data from sweep 3 was explored spatially. One of the most interesting findings from this work, is that whilst some interesting patterns emerge suggesting some differences in spatial distribution of young people who do graffiti or vandalism, the nature of these patterns is strongly affected by how the data are standardised. Just how data are standardised and presented via maps can have the potential to make a crime appear problematic or non-problematic dependent on the denominator chosen.

Figure 2 shows a large number of areas with unusually high levels or low levels of resident young people involved in vandalism per hectare (a value of 0.25 or below represents a quarter of the level that might be expected across Edinburgh, a value of 4 or above is 4 times the level expected for the city); correlation between vandalism and graffiti is very strong. Figure 3 uses the same data standardised by all respondents, thus it maps those involved in vandalism in an area compared to the level that might be expected out of all young people who live in the city. Here no areas stand out as having excessive levels of graffiti offenders; only one appears for vandalism. The correlation between

vandalism and graffiti is now medium.

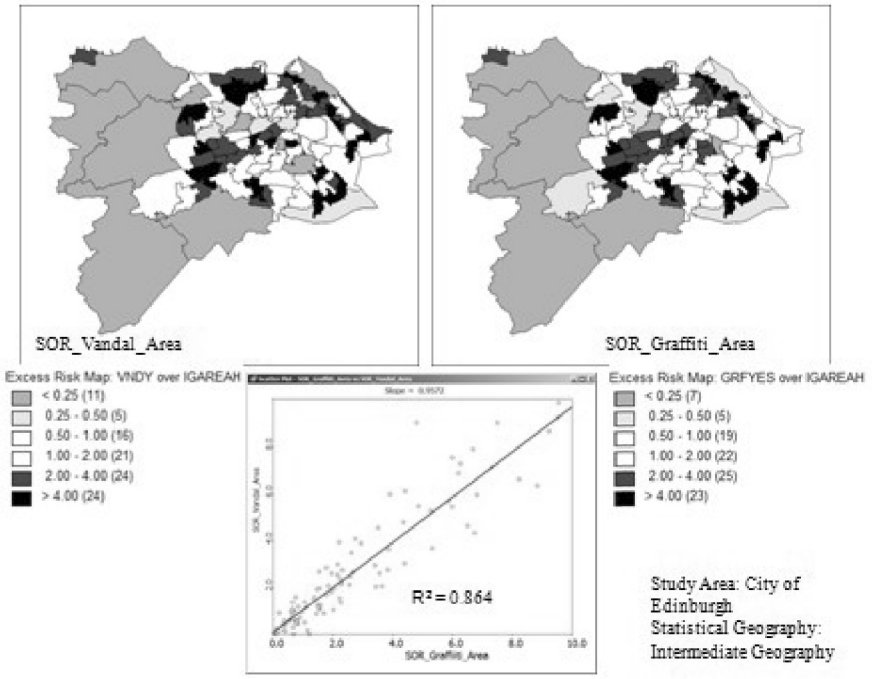



**Figure 2.** Young people who say yes to having done destructive vandalism (VNDY) or graffiti (GRFYES) per hectare Standarised Offence Ratios (SORs)

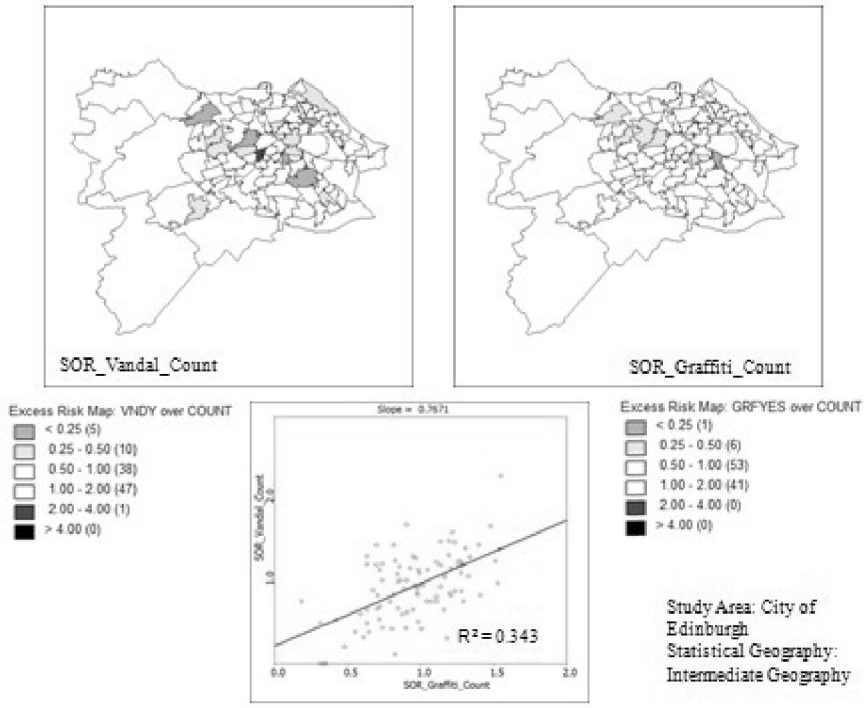



**Figure 3.** Young people who say yes to having done destructive vandalism (VNDY) or graffiti (GRFYES) per total respondents (COUNT) SORs

These maps and scatter plots produce very different results but with the same data. So which is right? Well the answer may be both. For a police officer working with local communities in Edinburgh Figure 2 might best represent the negative impact local residents' report of difficulties they face living with a high density of vandalism in their neighbourhood (assuming that young people offend where

they live). For a city youth worker Figure 3 reveals how vandalism appears to be a fairly normal behaviour for many young people, and Figure 2 just shows how some young people are singled out because there happen to be a lot of them living together in close quarters. Perhaps Figure 3 is more useful in that it identifies areas where vandalism is disproportionately high relative to the number of children living there.

This potential confusion is not resolved using different mapping break points (for example mapping by standard deviations from the mean, though it highlights different outliers depending on how the data are normalised).

Rather than visualising the geographical extent of a region, cartograms could be used to effectively convey the distribution of a population; enabling us to see the detail within, rather than between administrative boundaries. However, though cartograms are considered to be superior to choropleth mapping (eg Shaw et al 2008), they do not overcome the problems highlighted here. (Figure 4).

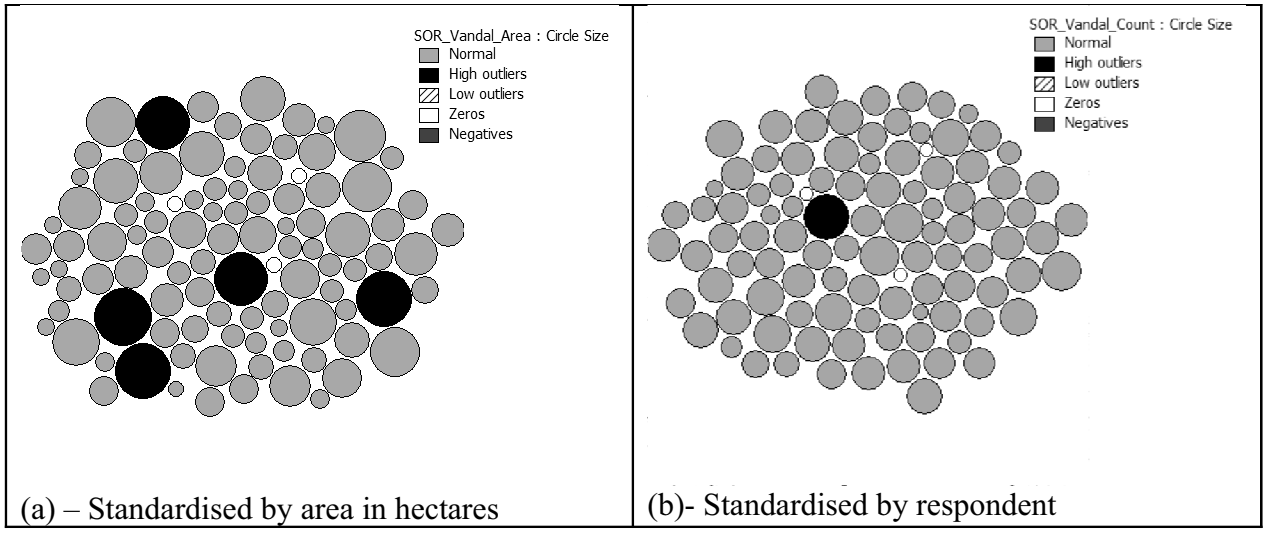



**Figure 4.** Young people who say yes to having done vandalism SOR per hectare (a) and per respondent (b) visualised as Cartograms

## 7. Conclusion

This work highlights generally that we need to think carefully about how we classify and standardise offences when we report them. It suggests some particular issues relating to how we should interpret and present crime data in maps. It illustrates that in interpreting crime statistics we need to be careful to understand how the methods by which the data has been standardised might affect results, and whether they are appropriate to the nature of the crime problem under consideration. Concern with how crime should best be normalised is not new but this "denominator dilemma" (Ratcliffe, 2010, 12) continues to be an issue (for example see Andressan, 2006); it is not yet being given sufficient consideration by academic and online mapping communities. With the growth of freely available government and local authority data (Rogers, 2010), and the introduction of public crime maps in England and Wales, there is an urgent need for further research and greater guidance on how to adequately represent the population at risk. There has been a particular lack of consideration of how we can best represent areas with persistent problematic vandalism, despite this remaining a common concern for many communities. As this project progresses it aims to provide further guidance.

## 8. Acknowledgements

Scottish Centre for Crime and Justice Research (SCCJR) funders of the PhD project. www.sccjr.ac.uk/. Susan McVie and Jackie Palmer, University of Edinburgh, Edinburgh Study Youth

Transitions in Crime. www.law.ed.ac.uk/cls/esytc/. Geooda and OpenGeoda have been developed by Luc Anselin (Anselin, 2005) and others and are available to download free at: geodacenter.asu.edu/

**References**

Anselin, L. (2005) *Exploring Spatial Data with Geoda: A workbook,* Urbana, Univeristy of Illinois.

Andressan (2006) Crime Measures and Spatial Analysis of Crime Activity, *British Journal of Criminology*, 46, pp258-285

Bottoms, A. E. (2007) Place space, crime and disorder In Maguire, M., Morgan, R. & Reiner, R. (Eds.) *Oxford Handbook of Criminology.* Oxford, Oxford University Press, pp528-574

Ceccato, V. & Haining, R. (2005) Assessing the geography of vandalism: Evidence from a Swedish city. *Urban Studies,* 42 pp1637-1656.

Chainey, S. & Ratcliffe, J. (2005) *GIS and Crime Mapping,* Chichester, Wiley.

Cohen, L. & Felson, M. (1979) Social Change and crime rate trends: a routine activities approach. *American Sociological Review,* 44 pp588-608.

Goldstein, A. P. (1996) *The Psychology of Vandalism*, New York, Plenum Press.

Levy-Leboyer, C. (Ed.) *Vandalism: Behaviour and motivations.* Amsterdam, Elsevier Science Publications

McVie S (2001) *The Edinburgh Study Of Youth Transitions And Crime:Technical Report Sweeps 1 and 2*, Edinburgh, University of Edinburgh

McVie S (2001) *The Edinburgh Study Of Youth Transitions And Crime:Technical Report Sweeps 3 and 4*, Edinburgh, University of Edinburgh

Melossi, D. (2008) *Controlling Crime Controlling Society Thinking about Crime in Europe and America*, Cambridge, Polity.

Ratcliffe, J. (2010) *Crime Mapping: Spatial and Temporal Challenges* in Piquero A. Weisburd D Handbook of Quantitative Criminology, New York, Springer

Rogers, S (2010) *Britain follows US with national data website Inventor of world wide web brought in for launch of data.gov.uk*, The Guardian, 21 January 2010. http://www.guardian.co.uk/technology/datablog/2010/jan/21/government-data-website-launched, accessed 15 February 2010

Shaw, M. Davey Smith, G, Thomas, B and Dorling: D. (2008) *The Grim Reaper's Road Map: An Atlas of Mortality in Britain,* Bristol, Health and Society Series: Policy Press

Smith, D. J. (2007) Crime and the life course In Maguire, M., Morgan, R. & Reiner, R. (Eds.) *Oxford Handbook of Criminology.* Oxford, Oxford University Press.

Smith D. J., McVie S, Woodward R, Shute J, Flint J & McAra L (2001) *The Edinburgh Study Of Youth Transitions And Crime: Key Findings At Ages 12 And 13*, Edinburgh, University of Edinburgh

**Biography**

*Ellie Bates is a second year PhD student in Criminology at the School of Law at the University of Edinburgh. Her interests include understanding vandalism, considering theories of crime and place, crime mapping, exploratory spatial analysis techniques and research communication.*

*William Mackaness is a senior lecturer in the School of GeoSciences, at The University of Edinburgh. His research interests lie in the modeling of geographic phenomenon at multiple scales, and visualisation methodologies.*

# Building Personalized Spatial Cognitive Road Network Based on Multivariate Binary Logistic Regression for Agent-Based Pedestrian Evacuation Behaviour Simulation

Lei Wu[1], Hui Lin[1]

[1] Institute of Earth and Space Information Science, The Chinese University of Hong Kong
Tel. (852)31634406
wulei@cuhk.edu.hk

KEYWORDS: Agent-Based Modeling, Personalized Spatial Cognitive Road Network, Evacuation Behaviour Simulation, Multivariate Binary Logistic Regression, Victoria Harbour Fireworks Display

## 1. Introduction

In the fields of pedestrian evacuation behaviour modelling, the Agent-Based Modeling paradigm has been applied to simulate the scenarios effectively includes researches done by Batty et al. (2003), Shi and Lin (2003), Kerridge et al. (2001), Bandini et al. (2006), and Ronald and Kirley (2006). Some of the crowd dynamics such as bottle-neck effect by Helbing et al. (2000), Hoogendoorn and Daamen (2005), and Zhao et al. (2006) were successfully analyzed and demonstrated. However among the above simulations, the action space of pedestrian agents was either a plain lattice board or a spatial road network while each cell or each road in the study area was assumed to be recognizable to each pedestrian agent. An assumption stood behind that the pedestrians were acquainted with the study area that they could take every cell or every road into consideration while making decision of selecting the routes for evacuation. However this assumption was illustrated to be somewhat unreliable by the field survey taken in a real fireworks display event in this paper. Through reviewing the survey data, the author found that different pedestrians could have different road network configurations and different roads could appear different significance to each pedestrian. Since being aware of the diversity of the spatial cognition, the author proposed the so-called Personalized Spatial Cognition Road Network (PSCRN) which was defined on the basis of individual spatial knowledge. By building the PSCRN, any two simulated pedestrian agents could probably find their own distinct routes to the same destination even though they were close to each other, which is quite ordinal in real life.

## 2. Data

In this paper, the fireworks display event at Victoria Harbour in Hong Kong was taken as the sample case. According to history records from Hong Kong Police Force H.K.P.F. (2008), about 300,000 tourists participated in each event, which is a sufficient large and diverse population for the study. The author made a questionnaire for the survey to investigate general background, spatial cognition,

response pattern, roads selecting inclination, routing strategy of each interviewee. Through this survey, the author aimed to find out the factors that affect the selection of the roads that made up the PSCRN while making decision their routes in case of evacuation, and build up the relationship between them quantitatively based on multivariate binary logistic regression.

The Sample case event took place from 7:00 pm to 9:00 pm on 1st Oct, 2008. Due to the tough field management pressure that Hong Kong police force sustained, we were allowed to distribute the survey forms in the study area from 3:00 pm to 6:00 pm. With the help of recruited student helpers, totally 680 survey samples were collected involving 1485 tourists of the event. Those who had companions were allowed to discuss to decide a choice to the question items collaboratively. Table 1 to Table 5 were a quick summarize of relative factors of the collected data.

**Table 1.** Gender

| Gender | *Percentage* |
|---|---|
| Male | 51% |
| Female | 49% |

**Table 2.** Age Group

| Age Group | *Percentage* |
|---|---|
| <18 | 16% |
| 18~30 | 49% |
| 30~60 | 32% |
| >60 | 3% |

**Table 3.** Education Level

| Education Level | *Percentage* |
|---|---|
| Secondary or Below | 48% |
| Bachelor | 36% |
| Master or Above | 18% |

**Table 4.** Residence Zone

| Residence Zone | *Percentage* |
|---|---|
| Within the district | 15% |
| Outside the district but in Hong Kong | 71% |
| Outside Hong Kong | 14% |

**Table 5.** Visiting Frequency(Per Month)

| Visiting Frequency | *Percentage* |
|---|---|
| Never | 13% |
| <4 | 50% |
| 4~10 | 24% |
| >10 | 13% |

## 3. Methodology

In order to establish the population sample of the frequencies of the roads being selected by some particular kind of pedestrian, an open question was set in the survey asking the interviewee to name the roads they would take in case of an emergency. The Figure 1 was the chart grading by the frequencies of the roads being selected offered as an overview.

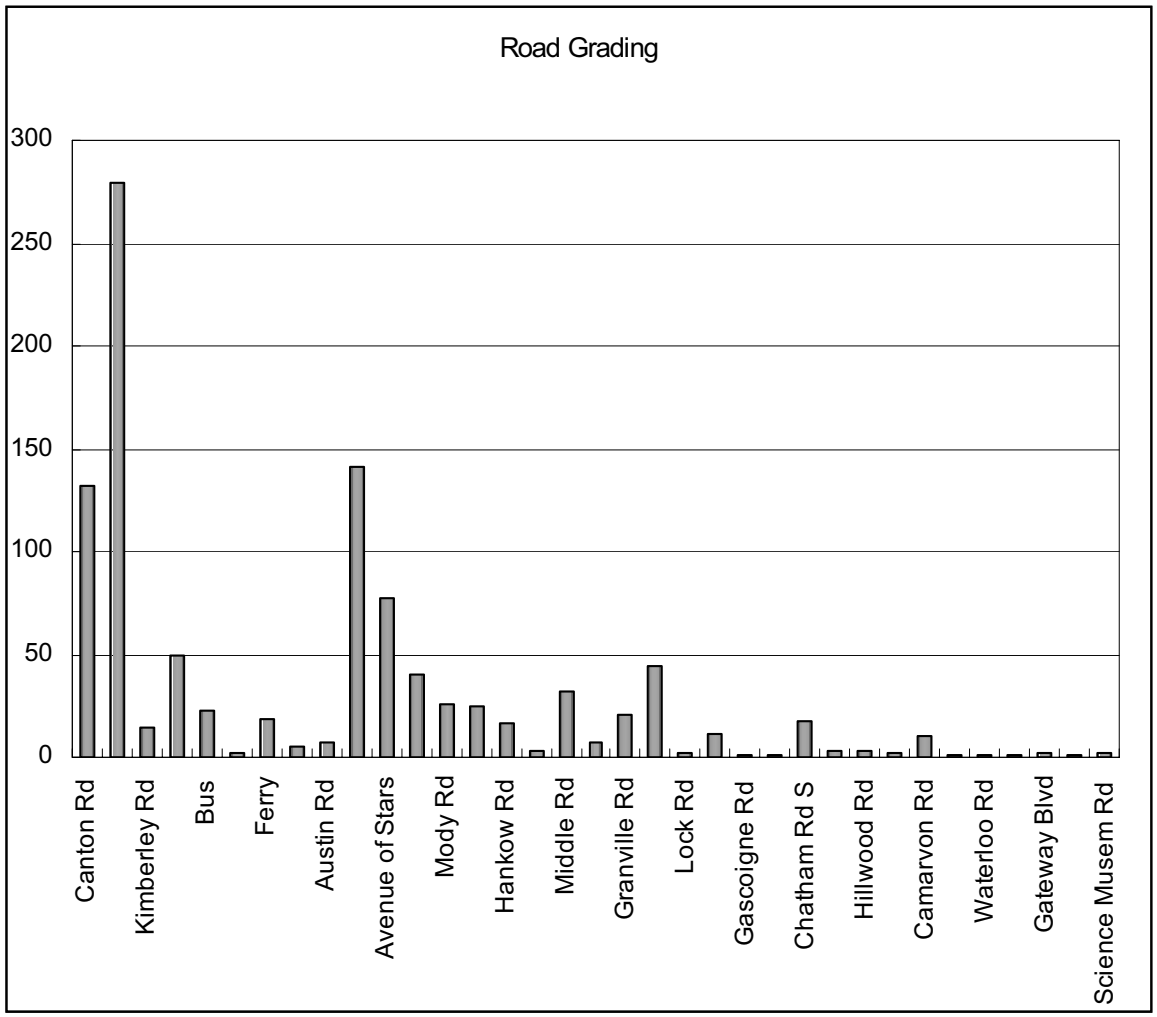



**Figure 1.** The Road Grading Chart

With the individual trial data, two steps were scheduled for building PSCRN to identify which roads that certain particular pedestrian would make use of for evacuation routing. Every single road appeared in the Figure 1 was investigated in each iteration. The first step was to determine whether the testing factor was significantly related to the variance of the probability the investigating road being selected to the pedestrian's PSCRN so that the testing factor should be included in the model through a Chi-Square test. The second step was then to establish the regression model between the result of the trial as the dependent variable and the included factors passed the Chi-Square test as the independent variables. Since the involved variables were all nominal scales, the multivariate binary regression

model would be applied to fit the data.

Taking road R for example, let's define the variable y indicating the road selecting result. Then y had two possible outcomes, either being or not being included in someone's PSCRN. If we code 1 for being included and 0 for being excluded, then the conditional probability of being included given someone's features is denoted by π(x) as shown in Equation 1.

$$\pi(X) = P(y = 1|X) \tag{1}$$

where X means the vector of feature variables.
This was a typical dichotomous dependent variable discussed by Hosmer and Lemeshow (1989) thus a multiple logistic regression model may fit the distribution given by Equation 2.

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2}$$

where g(x) was the logit transformation as shown in Equation 3.

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \tag{3}$$

The initial independent variables that enter the model in the first round were the interviewed features in the survey including G(Gender), A(Age Group), E(Education Level), Z(Residence Zone), and T(Visiting Times Per Month). Since all these variables were also categorical which means the actual numeric value of the variables were merely only labels of different levels without any special numeric meaning, a collection of dummy variables were built to detach each state of the variables. Taking Age Group as an example, three dummy variables would be built as Table 6 illustrated.

**Table 6.** Dummy Variables for Age Group

| Age Group | $DA_1$ | $DA_2$ | $DA_3$ |
|---|---|---|---|
| <18 | 0 | 0 | 0 |
| 18~30 | 1 | 0 | 0 |
| 30~60 | 0 | 1 | 0 |
| >60 | 0 | 0 | 1 |

Supposing $x_j$ has $k_j$ levels, then the $k_j$-1 dummy variables would be denoted by $D_{ju}$ where u=1, 2, …, $k_j$-1.Thus the Equation 2 would be extended as Equation 4.

$$g(x) = \beta_0 + \sum_{u=1}^{k_1-1} \beta_{1u} D_{1u} + \sum_{u=1}^{k_2-1} \beta_{2u} D_{2u} + \cdots + \sum_{u=1}^{k_g-1} \beta_{gu} D_{gu} \tag{4}$$

## 4. Model Building

With the general formula built, the author was able to establish the PSCRN for each pedestrian agent in the simulation by applying the regression to each road respectively substituting the agent features. In the simulation, once an agent was generated into the simulation, his personal features were decided randomly with respect to the population distribution of such features collected from survey. Then substituting these features to a particular road regression expression the probability that this road was included in his PSCRN could be calculated.

Here the Canton Road was taken as the example to demonstrate the building process. The probability that this road was selected into the PSCRN of someone agent was denoted by $\pi_{Canton}$, and the initial predictors were Gender, Age Group, Education Level, Residence Zone, and Visiting Frequency. In this paper, the author used SPSS(Version 17.0) to carry out statistics operations. In the first step to determine which factors were not related to selecting Canton Road by Chi-Square test, multiple two-way tables were constructed for testing correlations. Table 7 and Table 8 demonstrated the process for determining the gender factor.

**Table 7.** Cross Tables for Selection*Gender

| | | *Gender* | | |
|---|---|---|---|---|
| | | 1 | 2 | Total |
| Road Selection | No | 286 | 262 | 548 |
| | Yes | 60 | 72 | 132 |
| Total | | 346 | 334 | 680 |

*Gender 1 represents for Male, Gender 2 represents for Female

**Table 8.** Chi-Square test for Selection Gender

| | *Value* | *df* | *Asymp. Sig. (2-sided)* |
|---|---|---|---|
| Pearson Chi-Square | 1.931[a] | 1 | .165 |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 64.84

Here as the two-sided asymptotic significance value is greater than 0.1, it could be concluded that the selection probability of the Canton Road and Gender were unrelated which implies the factor of gender could be removed from the model. Table 9 summarized the significance testing over other factors.

**Table 9.** Pearson Chi-Square Values for Other Factors

| Factor | *Value* | *df* | *Asymp. Sig. (2-sided)* |
|---|---|---|---|
| Age Group | 8.150 | 3 | .043 |
| Education Level | 7.708 | 2 | .021 |
| Residence Zone | 1.971 | 2 | .373 |
| Visiting Frequency | 10.862 | 3 | .012 |

Therefore, for the Canton Road, the effective factors that could statistically significant predicting the probability of being selected into someone's PSCRN included Age Group, Education Level, and Visiting Frequency. Then the model was refined as Equation 5 shown where the categorical factors were expanded by dummy variables.

$$\pi(x) = \frac{e^{\beta_0+\beta_{11}DAge_1+\beta_{12}DAge_2+\beta_{13}DAge_3+\beta_{21}DEdu_1+\beta_{22}DEdu_2+\beta_{31}DTimes_1+\beta_{32}DTimes_2+\beta_{33}DTimes_3}}{1+e^{\beta_0+\beta_{11}DAge_1+\beta_{12}DAge_2+\beta_{13}DAge_3+\beta_{21}DEdu_1+\beta_{22}DEdu_2+\beta_{31}DTimes_1+\beta_{32}DTimes_2+\beta_{33}DTimes_3}} \quad (5)$$

In the second step to build linkage between the dependent variable of probability and independent

variables of related factors by the multivariate binary logistic regression, the method of independent variables entering the fitting model was set to be "Forward Selection (Likelihood Ratio)" so that the three factors which passed the Chi-Square test showing somewhat individually related to the probability were optimized again in a joint pattern. Table 10 gave the result of the regression.

**Table 10.** Regression Result for the Predicting Model

| | | *B* | *S.E.* | *Wald* | *df* | *Sig.* | *Exp(B)* |
|---|---|---|---|---|---|---|---|
| Step 1a | Timespermonth | | | 10.372 | 3 | .016 | |
| | Timespermonth(1) | -1.181 | .446 | 7.015 | 1 | .008 | .307 |
| | Timespermonth(2) | -.444 | .284 | 2.442 | 1 | .118 | .641 |
| | Timespermonth(3) | -.031 | .305 | .011 | 1 | .918 | .969 |
| | Constant | -1.083 | .247 | 19.291 | 1 | .000 | .338 |
| Step 2b | EducationLevel | | | 7.157 | 2 | .028 | |
| | EducationLevel(1) | -.668 | .268 | 6.240 | 1 | .012 | .513 |
| | EducationLevel(2) | -.252 | .269 | .877 | 1 | .349 | .777 |
| | Timespermonth | | | 9.938 | 3 | .019 | |
| | Timespermonth(1) | -1.184 | .448 | 6.992 | 1 | .008 | .306 |
| | Timespermonth(2) | -.444 | .286 | 2.405 | 1 | .121 | .642 |
| | Timespermonth(3) | -.054 | .307 | .031 | 1 | .861 | .948 |
| | Constant | -.688 | .317 | 4.699 | 1 | .030 | .503 |

a. Variable(s) entered on step 1: Timespermonth.

b. Variable(s) entered on step 2: EducationLevel.

Therefore the predicting model finally was in the form as shown in Equation 6.

$$\pi(x) = \frac{e^{-0.668\,DEdu_1 - 0.252\,DEdu_2 - 1.184\,DTimes_1 - 0.444\,DTimes_2 - 0.054\,DTimes_3}}{1 + e^{-0.668\,DEdu_1 - 0.252\,DEdu_2 - 1.184\,DTimes_1 - 0.444\,DTimes_2 - 0.054\,DTimes_3}} \tag{6}$$

And Table 11 gave the test of goodness-of-fit.

**Table 11.** Hosmer-Lemeshow Statistic Test

| Step | *Chi-square* | *df* | *Sig.* |
|---|---|---|---|
| 1 | .000 | 2 | 1.000 |
| 2 | 1.415 | 7 | .985 |

Since the final significance value of the test is 0.985, the predicting model fitted the data quite well. Thus it was indicated that the probability of the Canton Road being recognized by certain pedestrian could be estimated through substituting his education level and visiting frequency into Equation 6. And by repeating this computing flow, the regression models for each involved road could be established according to survey data.

## 5. Conclusion and Discussion

In this paper, the concept of PSCRN was proposed to represent the diversity of spatial cognition in the real life as exposed in the survey of the sample case. By applying the multivariate binary logistic regression to the categorical independent and dependent variables, the PSCRN was implemented that every road in the affected area had a unique probability distribution of being recognized by pedestrian agents according to their features. With the PSCRN model, two spatially adjacent pedestrians could probably have different solutions of preferred route to the same destination due to different configurations of PSCRN resulted from different personal features in agent based simulations, which could better represent the diversity of evacuation behaviour among a large amount population in the real world.

The sample case demonstrated the working flow of establishing the PSCRN model for the particular sample site. However when applying this model in other events, a field survey investigating the factors which could be used of reflecting the pedestrians' cognitions of spatial configurations in their minds is expected to be conducted yet. Another issue hasn't been solved is how to evaluate this model in evacuation simulations due to the lack of data collected in real emergencies. Possible attempts have been scheduled including comparing the population density generalized from the PSCRN model with measured data by the police force.

## 6. Acknowledgements

This paper was under the Competitive Earmarked Research Grant project titled as "Modeling Pedestrian Evacuation using a Geo-referenced Multi-agent Approach: Taking Hong Kong's Fireworks Show as a Sample Case" since August 2008 funded by the Research Grants Council (RGC) Hong Kong. The author greatly appreciated RGC supporting this research. And the author would like to express special thanks to the Hong Kong Police Force for permitting the survey conducting in the field, especially Mr. Andy Naylor and Ms. Amy Cheng of Tsim Sha Tsui Police Station of Hong Kong who offered many assistances on organizing and distributing the survey forms. Finally the author was very grateful to the student helpers from The Chinese University of Hong Kong for their hard working on operating the survey.

## References

Bandini S, Manzoni S and Vizzari G (2006) Toward a Platform for Multi-Layered Multi-Agent Situated System (Mmass)-Based Simulations: Focusing on Field Diffusion *Applied Artificial Intelligence* **20** pp327-351.

Batty M, Desyllas J and Duxbury E (2003) The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades *International Journal of Geographical Information Science* **17** pp673.

H.K.P.F. 2008. *A busy day for frontline officers* [Online]. Hong Kong Police Force. Available: http://www.police.gov.hk/offbeat/881/eng/n01.htm [Accessed].

Helbing D, Farkas I and Vicsek T (2000) Simulating dynamical features of escape panic

*Nature* **407** pp487-490.

Hoogendoorn S P and Daamen W (2005) Pedestrian Behavior at Bottlenecks *Transportation Science* **39** pp147-159.

Hosmer D W and Lemeshow S (1989). *Applied logistic regression.* Wiley, New York.

Kerridge J, Hine J and Wigan M (2001) Agent-based modelling of pedestrian movements: the questions that need to be asked and answered *Environment and Planning B-Planning & Design* **28** pp327-341.

Ronald N and Kirley M (2006) Pedestrian modelling: A comparative study using agent-based cellular automata *Computational Science - Iccs 2006, Pt 3, Proceedings* **3993** pp248-255.

Shi J and Lin H. Year. Simulating Pedestrian Evacuation Using Geographic Information Technologies. *In:* Proceedings of ACRS 2003 ISRS, 2003 Busan, Korea.

SPSS Inc.(2008) Helping Document Version 17.0

Zhao D L, Yang L Z and Li J (2006) Exit dynamics of occupant evacuation in an emergency *Physica a-Statistical Mechanics and Its Applications* **363** pp501-511.

**Biography**

*Lei Wu is a PhD candidate from the Institute of Earth and Space Information Science, The Chinese University of Hong Kong. His research interests including GIS applications in Public Management, Virtual Geographic Environments (VGE ), and Spatial Modeling and Simulation.*

*Prof. Hui Lin is the director of the Institute of Space and Earth Information Science, The Chinese University of Hong Kong. He is interested in the fields of Microwave Remote Sensing Image Processing and Analysis, Virtual Geographic Environments (VGE ), Spatial Database and Data Mining, and Spatially Integrated Humanities and Social Science.*

# Giving and Receiving Directions: Requirements for Automated Pedestrian Wayfinding Technology

Catherine Schroder, William Mackaness

The Institute of Geography, The University of Edinburgh, Drummond St, Edinburgh EH8 9XP
Tel: (44) 131 6508163 | Fax: (44) 131 6502524 | Email: c.j.schroder@sms.ed.ac.uk

KEYWORDS: Automated algorithms, landmarks, pedestrian navigation, route directions, wayfinding

## 1. Context

Imagine you are in the foyer of a hotel in a strange city; you have just arrived in town and are badly in need of a meal. You ask one of the staff members for a recommendation on where to eat. They suggest one of the current popular restaurants a few blocks away and give you directions on how to get there. But what kinds of information would you want included in these instructions? Getting from an origin A to a destination B is a recurring problem in everyday life and wayfinding is one approach to solving this problem. But how do you determine a route between A and B that ensures that 1) the individual is able to remember the instructions, 2) does not get lost, and 3) knows when they are lost? This research is particularly relevant to the design of automated wayfinding technologies. The focus is on the pedestrian.

The areas of wayfinding, route description, and the use of landmarks have received a lot of attention in research. The inclusion of landmarks or features of interest in route descriptions is central to this task. Previous research has shown that individuals refer more often to landmarks than to street names when generating route directions and they have been continuously proven to be more effective for pedestrians to follow than street names (Michon & Denis, 2001; Ross, May, & Thompson, 2004; Tom & Denis, 2003).

More recently research has moved on to look at ways in which landmark information could be automatically included in route descriptions (Caduff & Timpf, 2005; Elias, 2003; Nothegger, 2003; Richter, 2007). But questions remain over how to identify and automatically extract the most salient features of interest in the environment, or how to classify and thus prioritise their use in formulating route descriptions.

## 2. Method

In this paper we are primarily interested in the role features of interest play within route descriptions, in particular how they are used in giving directions and how they are described. How do these features ‘stand out’ – ie what makes them ‘salient’? The ambition is that the results of the following three experiments will be incorporated into a fully automated method of creating feature-rich route descriptions.

A set of three experiments were developed to examine these questions. The first experiment set out to identify the vocabulary used to describe features of the environment. This sought to identify the features in the environment that stood out (saliency) to the participant and to identify the reasons why they stood out, and what made them more salient than other features. The participants were walked along the route, in pairs, without knowing the destination and were only directed once they arrived at corners (for example ‘we will turn left here’). They were to identify and discuss any unusual, distinct, striking, or interesting features that stood out to them as they walked along the route.

In the second experiment we explored the way in which route descriptions are formed when traversing a route. This provided us with a very detailed set of descriptions that reduced down the large number

of features that were identified in experiment one and to start examining the reasons why certain features were selected more often than others and how they were included in the descriptions (primary direction cues, confirmatory cues, or 'you have gone too far' cues). The participants, in pairs, were led along a different route to the first experiment and again did not know the destination. They were asked to develop a set of route directions for an imaginary friend to follow as they walked along the route and were encourage to discuss the reasons behind their selection of each directional cue.

Finally, the third experiment looked at the recollection of the routes that were walked during experiments one and two. The participants were asked, individually, at the end of each experiment to provide a set of directions for the route that they had just walked from memory. This provided us with an insight into which features of interest the participant would choose if required to give a quick set of directions to someone but also enabled us to identify those features that were most memorable.

Two routes were used within the experiments (Figure 1). Experiment one and two were completed on separate routes whilst experiment three was completed at the end of the first two experiments. The two routes provided us with a way of looking at how the complexity of the route changed the information that was provided. This was especially important in experiment two and three. Additionally, each route was tested in both directions (for example A-B and B-A) as this allowed us measure the saliency according to directional differences in the information given along a route.

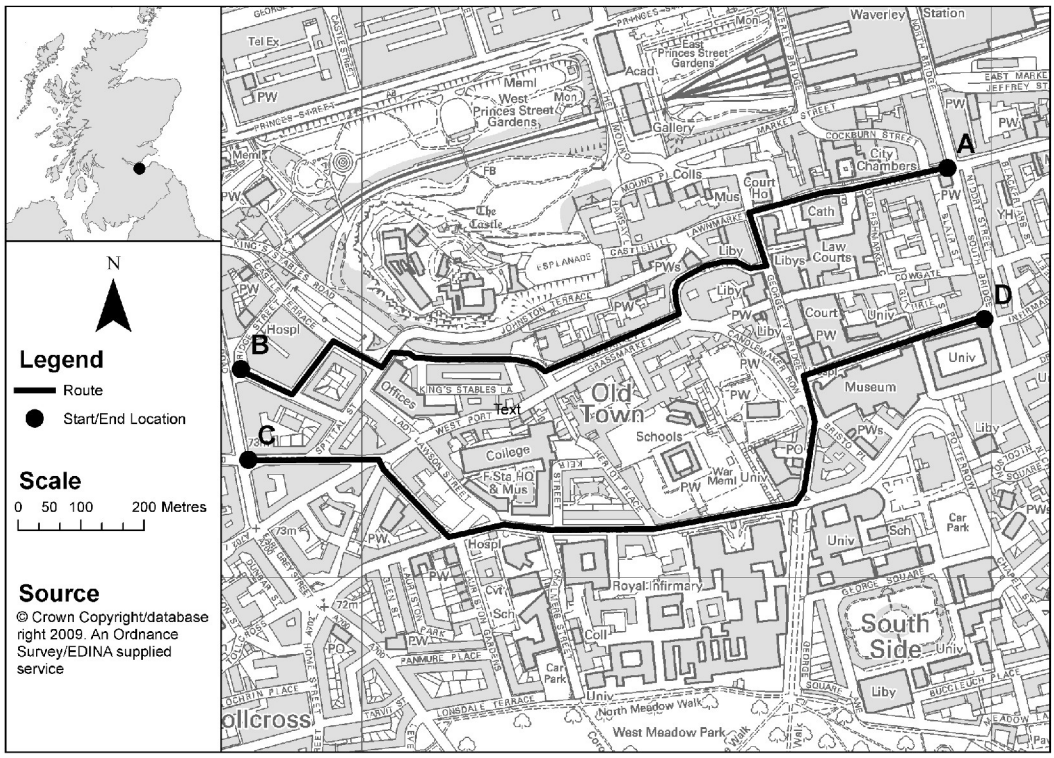



**Figure 1.** Illustration of the two routes: A-B and C-D. Route C-D is the simpler one.

Forty students from the University of Edinburgh, 20 males and 20 females, took part in the experiment, aged between 19 and 31. They had a large variation in the length of residency in Edinburgh (from 4 months to 11 years). The experiments were carried out in pairs to encourage open discussion and were recorded and videoed to help analysis.

## 3. Results

In the first experiment, we looked at the way in which the features of the environment were described. Many different descriptions were given, the most frequent related to the size, shape, and colour of the building. This confirms those variables used by Raubal and Winter (2002) in their landmark saliency model. However, a large number of other descriptors included the function of the feature (such as church or theatre), the age of the feature, architecture style, the façade construction, the gradient of the road and its shape.

In our second experiment we explored how route directions were given as people traversed the route. After simplifying the transcriptions into a minimal set of directions it was found that that the majority of the directions given (approximately 70% of each route) related to features of interest rather than streets. These directions were related to an action, an introduction, or a description of a feature. This confirms that features of interest are a more natural way to navigate by than street names. It was also found during this experiment that whilst a large number of features can be used to give directions, it was possible to identify a core set that were used by half or more of the participants. The participants traversed the two routes in both directions (eg. A-B and B-A) thus each route had two core sets of features. This is an important observation. It illustrates that saliency is relative to other features in the field of view, and their degree of visibility (which depends on the direction of approach). Therefore the relative differences between features need to be taken into account (for example buildings façade colour may differ). The difference in the surrounding area approaching the feature of interest on either side must be modelled in any automated system.

The third experiment found that whilst using a feature of interest to prescribe a direction action was the most frequent directional cue used, the second most frequent was prescribing a direction action using a street. This was in contrast to the previous experiment. This reflects that once away from the route the participants rely more on their memory, hence their prior knowledge of the study area is having an effect on their recall of the route. Additionally the street names that they were remembering were the more well known streets within Edinburgh, such as the Royal Mile and the Grassmarket. Finally, the majority of the directional cues that the participants gave in their route descriptions were based around the reorientation points of the route, whilst only a few used additional directions along the roads in a confirmatory ‘you are on the right track’ manner. Table 1 shows an example of the type of features and their descriptions used within the route directions for experiment three. Several of these features are illustrated in Figure 2.

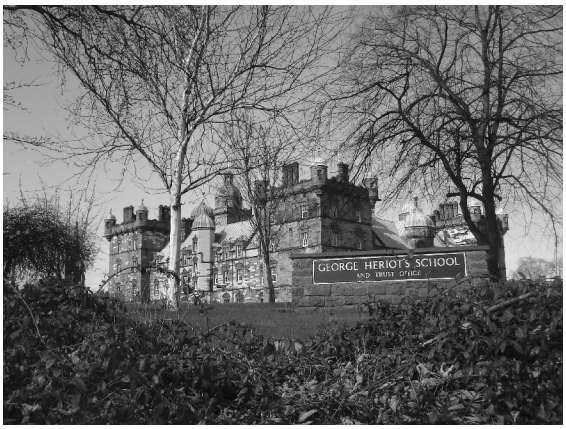
**Table 1.** Example route directions from experiment three, using route D to C

| *Description* |
|---|
| Walk up Chambers Street |
| Past the Royal Museum |
| Past the National Museum of Scotland |
| The National Museum is a round red building and quite new |
| Turn left |
| Face the church |
| The church is old looking and is dark with red doors |
| Stay on the right side of the church |
| Walk down the street |
| The street is short |
| Turn right |
| Continue straight |
| You will go past a school |
| The school is a big huge building and looks like a small castle |
| Continue straight for a long time |
| The street bends |

| |
|---|
| Veer to the left where the road bends |
| Turn right when you get to the Premier Inn on the right side |
| Walk until you come to the construction site |
| See the Nail Polish Bar |
| The Nail Bar is purple |
| Take the second road to the left |
| Continue straight |
| Keep walking until you reach the Odeon |

(a)

(b)




(c)

(d)

**Figure 2.** A selection of the landmarks used in the description in Table 1. (a) National Museum of Scotland. (b) Bedlam Theatre (converted church). (c) George Heriots School. (d) Construction site.

## 4. Conclusion

We contend that these three experiments provide the empirical evidence to support the modelling requirements in any automated route direction system and they help identify the criterion that governs the saliency of features of interest in the urban environment. It is essential to note that there are a number of ways that the saliency of a feature can be measured, but it must take into account the size, shape, and colour of the feature as well as its function. Finally, the most important part of the saliency of a feature of interest is the fact that it must stand out from the surrounding area. It is therefore essential to develop a measure of difference to select the most salient feature to use at each

possible decision point. This difference measure must take into account the visibility from the decision point, which will be different depending on the direction of travel.

## References

Caduff, D., & Timpf, S. (2005). The Landmark Spider: Representing Landmark Knowledge for Wayfinding Tasks. In T. Barkwosky, C. Freksa, M. Hegarty & R. Lowe (Eds.), *AAAI 2005 Spring Symposium*. Stanford, CA: AAAI Press.

Elias, B. (2003). Extracting Landmarks with Data Mining Methods. In W. Kuhn, M. F. Worboys & S. Timpf (Eds.), *COSIT 2003* (pp. 375-389). Berlin Heidelberg: Springer-Verlag.

Michon, P. E., & Denis, M. (2001). When and Why are Visual Landmarks Used in Giving Directions? In D. Montello (Ed.), *Spatial Information Theory* (pp. 292-305). Berlin: Springer.

Nothegger, C. (2003). *Automatic Selection of Landmarks.* University of Technology, Vienna.

Raubal, M., & Winter, S. (2002). Enriching Wayfinding Instructions with Local Landmarks. In M. J. Egenhofer & D. M. Mark (Eds.), *Geographic Information Science* (pp. 243 - 259). Berlin: Springer.

Richter, K. F. (2007). A Uniform Handling of Different Landmark Types in Route Directions. In S. Winter, M. Duckham, L. Kulik & K. B. (Eds.), *Spatial Information Theory* (pp. 373-389). Berlin: Springer-Verlag.

Ross, T., May, A., & Thompson, S. (2004). The Use of Landmarks in Pedestrian Navigation Instructions and the Effects of Context. In S. Brewster & M. Dunlop (Eds.), *Mobile HCI 2004* (pp. 300-304). Berlin Heidelberg: Springer-Verlag.

Tom, A., & Denis, M. (2003). Referring to Landmark or Street Information in Route Directions: What Difference Does it Make? In W. Kuhn, M. F. Worboys & S. Timpf (Eds.), *COSIT 2003* (pp. 362-374). Berlin Heidelberg: Springer-Verlag.

## Biography

*Catherine Schroder is a PhD at The University of Edinburgh. William Mackaness is a senior lecturer in the School of GeoSciences at The University of Edinburgh.*

# Cross-Scale Movement Trajectory Analysis

Patrick Laube[1], Ross Purves[2]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
Tel. +41 44 635 65 34
Email: patrick.laube@geo.uzh.ch, www.geo.uzh.ch/~plaube

[2]Tel. +41 44 635 65 31, ross.purves@geo.uzh.ch, www.geo.uzh.ch/~rsp

KEYWORDS: Movement analysis, cross-scale analysis, sampling, trajectories, GPS-tracked animals.

## 1. Introduction

Individual movement can be characterised by descriptive movement parameters such as speed, sinuosity, or turning angle. Such movement parameters are normally derived from *trajectories*, that is two-dimensional time-stamped poly-lines. However, just as with the classic example of 'slope' derived from a terrain in geomorphometry, there is no *true* speed or sinuosity for a given time-stamp on a trajectory since any such attribute is defined as a function of sample point spacing, – or *scale* (Goodchild, 2001).

In this paper we aim to explore the sensitivity of a range of derived parameters to temporal scale, largely motivated by our access to a rather unique data set, featuring grazing cows whose position was recorded with very high temporal resolution. Such data allows cross-scale analysis of movement properties, enabling sampling at sub-second, minute, or even hourly scales. In this experiment we therefore address the specific research question – to what degree do movement parameters such as speed, sinuosity, or turning angle vary when derived at variable temporal scales?

We present a method where movement parameters are computed along trajectories, whilst the temporal analysis scale is systematically varied. The statistical properties of the resulting scale classes are visualised in box whisker plots, mapping speed, sinuosity, and turning angle as a function of temporal analysis scale. Initial experiments with two trajectories capturing the movement of cows in a paddock illustrate our method.

## 2. Background

In ecology, it is widely acknowledged that the key to the understanding of observed phenomena lies in the elucidation of the mechanisms intertwining pattern and scale (Levin, 1992). In GIScience, cross-scale methods investigating a geographic phenomenon at multiple scales have, for example, been exemplified for the measurement of the shape of the earth's surface (Fisher *et al.*, 2004) and point clustering (Lu and Thill, 2008). Whereas the former reference stands for the crucial discussion on the scale-dependant definition of vague spatial features, the latter represents research identifying critical scales at which abrupt changes of patterns occur, or invariance across scales. Both examples illustrate that any selection of scale for modelling or analysing spatial phenomena may significantly influence what we see and hence how we understand the phenomenon under study.

When researching methods for cross-scale analysis for movement data, a side trip beyond GIScience and into behavioural ecology is rewarding. For decades, biologist have recorded and analysed the paths of animals of all sizes and across various scales. For example, Fryxell et al. (2008) summarise several studies investigating animal movement at three different scales (coarse-scale, intermediate-scale, fine-scale). However, most such cross-scale studies focus on biology and not methods, and furthermore rely on different data sources for different scales and hence hardly allow detailed

methodological cross-scale studies.

Nams (2005), however, is an exception relevant to this experiment as he explicitly derives fractal dimension *D* (as a measure for sinuosity, or *tortuousity* as it is often termed in behavioural ecology) for the same trajectories at various scales. The hypothesis underlying his research states that animals express different movement behaviours (e.g., tortuous foraging at fine scales but directed advances at coarse scales) at different scale sections (so called 'domains'), which are identified through cross-scale analysis. In contrast to Nams' work focussing on space-only fractal dimension *D*, we suggest (a) the joint analysis of several movement parameters (speed, turning angle, *and* sinuosity), (b) the inclusion of inherently spatio-temporal movement parameters (e.g. speed), and (c) the explicit analysis of the measurement variance to be found at different analysis scales through box whisker plots.

## 3. Data and data preprocessing

The trajectories emerged from a smart farming project applying wireless sensor technology, carried out with CSIRO ICT Centre and CSIRO Livestock Industries, Rockhampton, Australia. Ten cows were tagged with GPS receivers and their trajectories were monitored whilst the cows were grazing on a paddock of 600m*200m at the CSIRO Rockhampton research site. The data cover roughly three days of continuous tracking at a sampling rate of four fixes per second. Out of the ten individuals two focus trajectories were selected for this experiment (#1008 and #1016), each featuring approximately one million fixes. For this experiment phases of extended resting were excluded by segmenting trajectories with resting phases (not moving beyond a threshold $d$ = 0.6m for 10 minutes) serving as segment separators (45 segments for #1008, 26 for #1016).

## 4. Methods

In this paper we explore both *temporal sampling intervals* (the effects of the underling temporal resolution of our data) and the *temporal analysis scale* at which we measure a parameter (the measurement window over which we calculate a value). For the initial experiments presented in this abstract, the width of the moving measurement window was set to the same temporal analysis scale as the sampling interval (that is for sampling interval $s$ = 10sec a window of width $w$ = 10 sec). Hence, in the remainder of this paper the term 'scale' refers both to temporal sampling interval and temporal analysis scale. The sampling/analysis ratio, however, need not be constant but is a parameter of the presented method and its influence shall be investigated in further work.

The method investigates movement parameters for a series of temporal scales (0.5sec, 1sec, 10sec, 1min, 2min, 5min, 10min, 0.5h, 1h, 2h, 3h). At every given temporal scale, a moving window was shifted along the trajectory, systematically sampling the investigated movement parameter (Laube *et al.*, 2007). See Figure 1 for an illustration with three temporal scales $s_1$, $s_2$, and $s_3$ and three sampling windows $w_1$, $w_2$, and $w_3$. Note that sampling was only permitted within continuous segments between resting phases. All sampled values per scale were binned and each bin resulted in an item on the ordinate of the box whisker plots (see Figure 2).

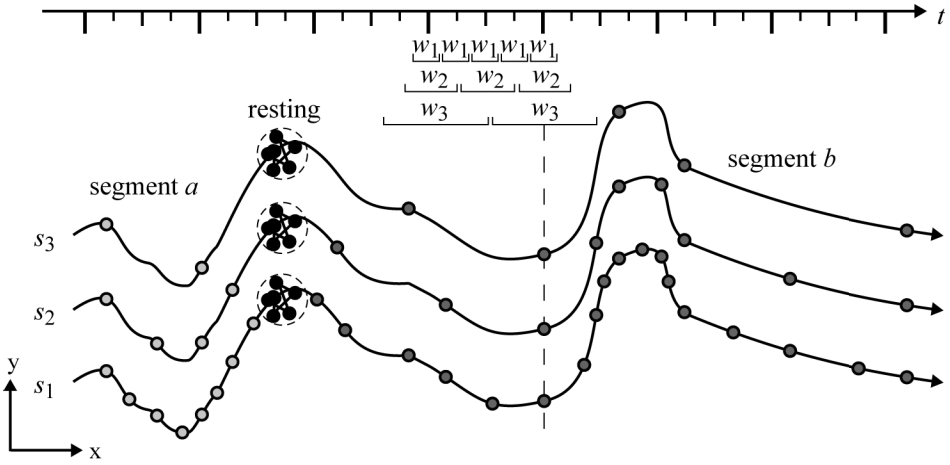



**Figure 1.** Deriving movement parameters at variable temporal scales.

Given a window *w*, *speed* and *turning angle* were computed for every *fix(t)* as the centered average for three fixes, sampled at $t_{-0.5w}$, $t$, and $t_{+0.5w}$. Although there exists a plethora of ways to term and compute 'wigglyness' of a poly-line (Claussen *et al.*, 1997), we chose *sinuosity* (Dutton, 1999), as the ratio between the actual track length and the line connecting the end points of the sampling window *w* (values of 1 representing straight trajectories, values > 1 indicating increasingly sinuous trajectories).

A Java application was coded for computing the trajectory parameters, and the R statistics environment was used for statistics and the box whisker plots. The box whisker plots show medians (horizontal bar), 25th and 75th percentiles enclosing the middle 50% of the data (boxes, also interquartile range, IQR), minimum and maximum values (whiskers), and outliers (data points more than 1.5 times the IQR from either end of the box).

## 5. Preliminary results

Figure 2 illustrates multi-scale box-whisker plots for speed, sinuosity, and turning angle for the two cows. Note that both speed and sinuosity plots do not show all values and that the sinuosity plot has a logarithmic y-axis. As a first observation both plot series are strikingly similar. This finding can be explained by the fact that both individuals move within a herd for most of the time, showing significant movement coordination.

Very short scale intervals measure the instantaneous magnitude of velocity, since sinuosity at these scales is ~1. As sinuosity increases with scale, speed typically decreases since the actual displacement resulting from a sinuous path is less than the path length. *Sinuosity* and its variance increase with scale, levelling out above a factor 10 at a scale of about two minutes. Perhaps surprisingly the two finest scales show very little variance, as we might expect considerable noise for narrow windows. However, these scales are very close to the raw data sampling of 4 fixes per second, and the signal to noise ratio is low after we filter resting cows, and explore movement parameters for moving cows. We speculate that the increase of sinuosity after 2h and 3h is probably due to the extent of the paddock. Grazing cows cross the paddock in around that time and are then pushed back by the fence, resulting in convoluted trajectories. Little variance is also found for *turning angle* at fine scales for the same reasons, though turning angles seems to contain little scale dependence in general.

A key result is that the joint analysis of speed and sinuosity reveals at which temporal scale speed is best measured. The low variance in sinuosity for the two finest scales suggests that cows diverge little from direct paths. That in turn implies that we can estimate the actual grazing speed of cows to be around $0.2 ms^{-1}$, which could then be used, for example, to estimate energy use over time.

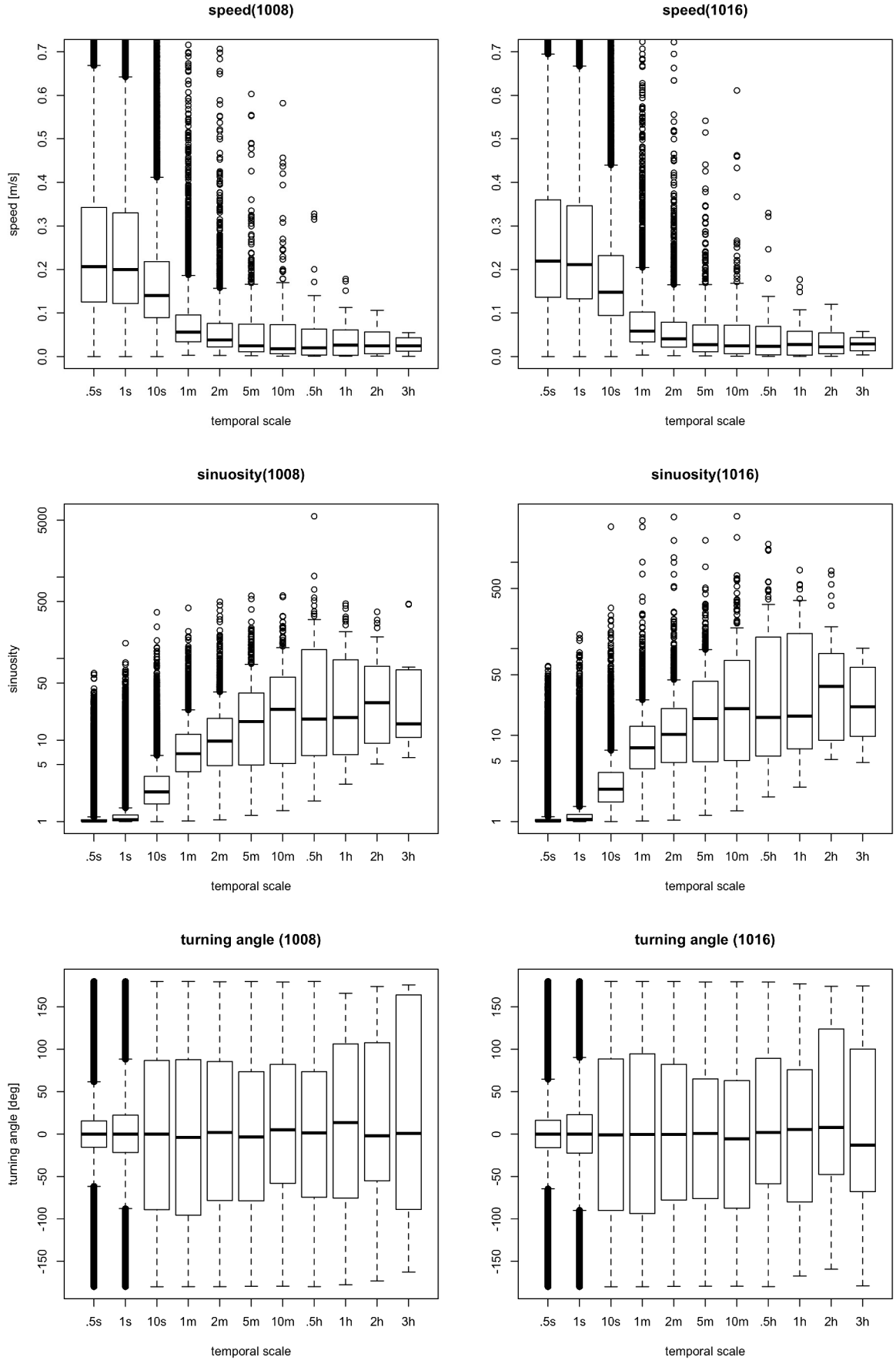



**Figure 2.** Cross-scale analysis of speed, sinuosity, and turning angle for two cows (#1008,

#1016).

## 6. Acknowledgements

The authors thank Tim Wark, CSIRO ICT Centre and CSIRO Livestock Industries, Rockhampton, Australia, for providing the excellent cattle data.

## References

Claussen D L, Finkler M S and Smith M M (1997) Thread trailing of turtles: methods for evaluating spatial movements and pathway structure *Can. J. Zool.-Rev. Can. Zool.* 75 pp2120-2128

Dutton G (1999) Scale, sinuosity and point selection in digital line generalisation *Cartography and Geographic Information Systems* 26 pp33-53

Fisher P, Wood J and Cheng T (2004) Where is Helvellyn? Fuzziness of multi-scale landscape morphometry *Transactions of the Institute of British Geographers* 29 pp106-128

Fryxell J M, Hazell M, Börger L, Dalziel B D, Haydon D T, Morales J M, McIntosh T and Rosatte R C (2008) Multiple movement modes by large herbivores at multiple spatiotemporal scales *Proceedings of the National Academy of Sciences* 105 pp19114-19119

Goodchild M F (2001) *Modelling Scale in Geographical Information Science,* ed N J Tate and P M Atkinson New York, NY: John Wiley and Sons pp 3-10

Laube P, Dennis T, Walker M and Forer P (2007) Movement Beyond the Snapshot - Dynamic Analysis of Geospatial Lifelines *Computers, Environment and Urban Systems* 31 pp481-501

Levin S A (1992) The Problem of Pattern and Scale in Ecology *Ecology* 73 pp1943-1967

Lu Y and Thill J-C (2008) Cross-scale analysis of cluster correspondence using different operational neighborhoods *Journal of Geographical Systems* 10 pp241-261

Nams V O (2005) Using animal movement paths to measure response to spatial scale *Oecologia* 143 pp179-188

## Biographies

*Patrick and Ross are both lecturers with the Department of Geography, University of Zurich. Patrick's main research interests are movement analysis, spatio-temporal data mining, and most recently decentralised spatial computing for geosensor networks. Ross' research interests include geographical information retrieval, geomorphometry, and uncertainty.*

# SIMULACRA[1]: A New Land Use-Transportation Modelling Framework for London

Michael Batty[+]

[+]CASA, UCL, 1-19 Torrington Place, London WC1E 6BT, UK
Tel. (44) 207 679 1781
Email: m.batty@ucl.ac.uk | www.casa.ucl.ac.uk

KEYWORDS: land use transportation, spatial interaction, housing markets, energy costs, behavioural predictions

## 1. Introduction

We are building an integrated, quasi-dynamic model for Greater London and the Thames Gateway as part of an effort to develop models for simulating the impact of changes in spatial behaviours over relatively long time periods of up to 50 years. This model is part of our interest in assessing the impact of dramatic shifts in the costs of energy related to travel and residential development on the location of population and employment. Such changes are part and parcel of the effects of climate change, aging and resources depletion and substitution in the energy sector which are forecast to have a significant impact on location and spatial interaction patterns during the rest of this century. The model as it is currently designed is a cross-sectional, hence static, residential location model incorporating four modes of transport which reflect the cost of travel and the income budgets of populations with respect to wages, house prices and travel costs. The model is currently in pilot form for Greater London which is comprised of 633 spatial units (wards) and is structured so that it can be driven visually by an expert user from the desktop. It has links to external web-based visualization software so that it can be demonstrated effectively to stakeholders and it is currently part of the integrated assessment of climate change in the London region being developed in the Tyndall Centre for Climate Change Cities project (Dawson, et al. 2009) and its successor ARCADIA[2]. In the paper presentation, we will outline the mathematical structure of the model, develop its algorithmic basis, explore the visual basis to the simulation, and illustrate how it can be used to enable users to make predictions of future residential population in the London region, on-the-fly so-to-speak. The current version is a pilot and it is intended to develop at least two other versions of a much more comprehensive quasi-dynamic structure for a wider area than Greater London, comprising some 2000 or so zones in the metro south-east of England, embedding these versions into the London database that we are constructing, and relating this to various web-based services for visualisation.

## 2. Formal Structure of the Model

The model essentially distributes employment from place of work to residential locations over four transport modes, which compete for their share of travellers with respect to the average budget that each employee has at their place of work for travel. At each place of work, we know the average income of the employees and thus it is possible to compute an average disposable income available for transport. The same can be done for housing in that we can compute an average spend on housing for each employee and just as the average spend on transport is matched against the travel cost on each mode, the disposable income for housing is matched against house price in each residential zone. These functions are represented as deviations from the ideal budget and thus as an employee deviates

[1] **SIMULACRA** is our tentative acronym for the suite of London models we are building: **SIMU**lating **LA**nd use, **C**ommercial and **R**esidential **A**ctivities

[2] http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/G060983/1

further from the actual amount they wish to spend on transport, the probability of their travelling on that mode and to that residential locations decreases. The same structure is used for differences between their housing budgets and house prices. These various functions are reconciled in the usual spatial interaction accounting framework, which is used to constrain spatial models. In fact, the model is also subject to constraints on residential population, which is turn, are computed as a function of land use availability and density. In short, the model is an origin-constrained, semi-destination constrained spatial interaction model where costs are incorporated as deviations from the average with respect to transport and house prices. The algebraic formulation of the model is generated using entropy maximising which shows that the model is constrained to reflect average dispersion around house prices at each location and travel costs on each link, with a structure similar to that outlined in the probabilistic residential location models outline by Batty (1976, Chapter 10, pp. 264-270).This model structure is being extended to deal with other sectors of the urban system as in Lowry type models (Batty, 2009) but currently we only model the residential sector as in earlier versions of this style of model that have been linked to GIS and visualisation (Batty, 1992).

## 3. Visual Simulation

The model is programmed to run in visual desktop environment. It is built around the three classic stages of model-building: data exploration, calibration, and prediction. At each of these stages, the system involves data that needs to be displayed and explored visually in that the complexity of such applications –633 zones and the square of this for the number of trips on each of four networks – require visual display so that both the model builder and the user can grapple with the logic of what the data is implying and what the model is saying about this. This is not a matter of embodying the model in GIS – it could be built in **ArcGIS** although the overhead is substantial – but map displays are needed at every stage and transformations of these into various sort of cartogram are helpful. These kinds of models treat transport in great detail and require good flow mapping. As such the model user in this system is able to load the data and explore it on the screen. This is essential too as with so many zones errors can creep in as they have done in this application and the only way to detect these is visually. At each of the three stages - data, calibration and prediction - the user has access to maps of any of the input and output data, flow maps, tree maps and plots of trip distributions. We also have a facility to let the user display the data input or output in **Google Earth** which can be loaded within the model – while the model is running – so that users can ground the data and outputs in the model with other standard data available in **Google Earth**. It is even possible in this context to display the model data which is abstract with more realistic data such as the three content that we have for London from our Virtual London model. We consider that this kind of visual simulation is essential for complex models such as these (Batty, 1992).

## 4. Behavioural Forecasting: Sketching Future Scenarios for the Distribution of Population

After the model has been calibrated, we have the facility for exploring scenarios on-the-fly. Users can change any of the input data – in particular, sources and volumes of employment, travel costs on any of the networks through either the addition or deletion of new network links, through changes in income to be spent on transport which might reflect energy costs, through prices of transport and so on. We can also change land use attractions and constraints. So far, we cannot change population location except through constraints but in future versions of the model where we begin to model additional sectors, this will be possible. Users can thus build up scenario packages on-the-fly and immediately run the model making direct comparisons with the baseline input data and the calibrations. In this sense, the model can be used to explore the scenario space although the way to do this must be well structured if users are to converge on informed futures that are relevant to the general quest of examining these kinds of future over the long time scales involved.

Currently a relatively constrained class of scenarios has been examined in the Tyndall project (Hall, et al, 2009) and these have been used in a London stakeholder group comprising representatives from GLA, TfL EA, and related bodies. The planning support system we have fashioned is relatively robust and informal and in the new project ARCADIA more attention will be paid to how this science

can be most effectively used in scenario planning and forecasting through new methods of decision support.

## 5. Acknowledgements

Thanks to Steve Evans for providing the original data, to Duncan Smith for new data from the London database, and to Richard Milton who helped on the interface between the desktop model and visualisations in Google Earth. Stuart Barr and Ali Ford of Newcastle University generated the travel costs matrices for the baseline model.

## 6. References

Batty M (1976) *Urban Modelling: Algorithms, Calibrations, Predictions* Cambridge University Press, London

Batty M (1992) Urban modeling in computer-graphic and geographic information system environments *Environment and Planning B* 19 pp663-685.

Batty M (2009) Urban modeling, in R Kitchin and N Thrift (Editors) *International Encyclopedia of Human Geography* 12 Elsevier, Oxford, pp51–58.

Dawson R, Hall J, Barr S, Batty M, Bristow A, Carney S, Dagoumas A, Evans S, Ford, A, Harwatt H, Kohler J, Tight M, Walsh C, and Zanni A (2009) A Blueprint for the Integrated Assessment of Climate Change in Cities, in K. Tang (Editor) *Green CITYnomics: The Urban War Against Climate Change.* Greenleaf Publishing, Chippenham, UK, pp32-51.

Hall J W et al (2009) *Engineering Cities: How Can Cities Grow Whilst Reducing Emissions and Vulnerability* Tyndall Centre for Climate Change, Newcastle University, Newcastle, UK.

**Biography**

*Michael Batty is Bartlett Professor of Planning at University College London where he directs the Centre for Advanced Spatial Analysis (CASA). His research focuses on computer applications in urban analysis and planning and the work of his group can be seen at www.casa.ucl.ac.uk . He has was awarded a CBE in the Birthday Honours List in 2004 and is a Fellow of the Royal Society (FRS).*

# Spatially Clustered Associations in Health GIS "mashups"

Didier G. Leibovici[1], Lucy Bastin[2], Suchith Anand[1],
Jerry Swan[1], Gobe Hobona[1] and Mike Jackson[1]

[1]Centre for Geospatial Science, University of Nottingham,
Innovation Park, Triumph Road, NG7 2TU Nottingham, UK
Tel. +44 - (0)115 84 32760
didier.leibovici@nottingham.ac.uk, www.nottingham.ac.uk/cgs
[2]School of Engineering and Applied Science, Aston University, UK

KEYWORDS: spatial clustering, multivariate associations, co-occurrences, risk factors, Health GIS

## 1. Introduction

Developments in standards of the Open Geospatial Consortium (OGC) and International Organisation for Standardisation (ISO) along with the server-side and client-side software to allow the implementation of geospatial web services, enable GIS "*mashups*" to be seamlessly assembled by combining datasets from various sources and semantics. These geospatial "*mashups*" have huge potential in the health and epidemiological context, to derive intelligent outcomes, such as disease mapping or clustering, environmental risk factor analysis, exposure analysis, orforecasting and modelling of epidemics. However, practical application of these techniques requires efficient geoprocessing services that use pertinent statistical methods or algorithms; and there is frequently a dilemma in balancing the pertinence of the spatial methodology and the efficiency of the web service in terms of performance. This tradeoff between methodological complexity and real-time performance is amplified by the many and complex data sources which are available to be added to a '*mashup'*, and emphasises the need for simple exploratory methods which allow multivariate analysis of spatial data.

Cluster detection among labelled spatial features (in this instance, cases of disease) has a long history in epidemiology, ecology and geography (Lawson et al.,2006). Myriads of tests have been proposed to allow testing for spatial clustering or testing the *location* of clusters (e.g., Kulldorff , 2003, 2006). For locating of clusters, the Besag and Newell (1991) test, the GAM from Openshaw et al. (1987), and the spatial scan devised by Kulldorff and Nagarwalla (1995) are the most well-known approaches. While some principles of these methodologies differ (Besag and Newell is particularly distinct from the other two "scan" methodologies), they all make use of two populations: the cases, and the population at risk, (Waller and Gotway, 2004). For example, the spatial scan statistic implemented in SaTScan (Kulldorf 1997) allows use of a discrete Poisson model to handle 'case-only' data, but only in the presence of underlying population data for the region under consideration. Thus 'non-case' data is implicitly constructed to correct for any driving spatial heterogeneity in the population at risk.
With less control on the sampling design and/or the underlying population, disease mapping and disease clustering can be difficult, due to the heterogeneity of the overall background population and /or population at risk (*e.g.* if you are interested in mapping the incidence of a disease affecting only children under the age of five). Using the wrong population at risk, or working without 'non-case' data can easily result in misleading estimations of significance or pattern in a disease mapping application. Therefore "*mashups*" of case data and their potential risk factors need to focus on the more appropriate problem of locating spatial *associations*.
The aim of this paper is to present a generic approach for the challenge of detecting clusters of bivariate or multivariate associations between attributes of one or more populations or spatial features.

## 2. Multiway exploratory omnibus detection

Multinomial cluster detection can be seen as a simpler type of multivariate association cluster detection, if one considers each category as a realisation of a point process. This situation is shown in Figure 1, where categories do not seem to cluster themselves, apart from the **star** category which could show a cluster on the west "corner". Looking at the unmarked point pattern, one sees two or three clusters, but definitely *one sees only one cluster of association* of **star**, **dot** and **square** in the lower east "corner".

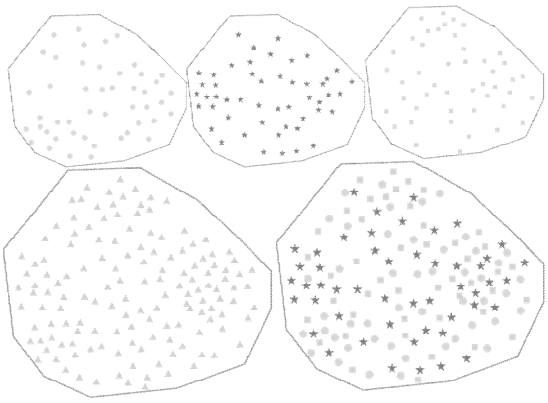


**Figure 1.** Hypothetical dataset of occurrences of three categories (top row): unmarked point pattern (bottom left) and marked point pattern (bottom right).

Recently Leibovici et al. (2008, 2009) developed an approach based on multiway contingency table co-occurrences of order ***k*** (***k>2***) to propose some exploratory methods allowing multinomial spatial dependence analysis. The CAkOO method uses a generalisation of correspondence analysis, (Leibovici, 2009), to decompose the chi-square of independence built from the multiway table whilst the SOOk method plots the entropy based on the multiway multinomial distribution of co-occurrences (for a chosen order) at different distances of collocation. CAkOO describes the spatial associations of categorical variables that are described without locating them, though some types of analysis allow spatial components to be displayed as well. SOOk, in a similar way to plotting a Ripley's ***K*** statistic, (Bivand et al.,2008), provides information on the spatial structuring of the co-occurrences at different scales (distances of co-occurrence).
These two methods are appropriate for overall detection of clustering / structuring by focusing on the problem of multiway and/or multivariate associations. However, they do not lend themselves directly to a delineated visualisation of association. Therefore, the declaration about Figure 1 -*one sees only one cluster of association*- needs a spatial scan approach to be fully assessed.

## 3. Clustered association detection,

The hypothetical data of Figure 1 could correspond to the identification of a multi-factorial zone of contagion (each point being a case and each category identifying a factor of contagion). Often the same point will carry more than one attribute (a multivariate point process) and multivariate multinomial co-occurrences analysis can identify profile clustering. The proposed method, called ScankOO, is inspired by the above-mentioned, widely-used cluster detection methods and exploits the spatial pattern that can be identified in high-order co-occurrences. The goal here is to build a statistical map that can be tested, using either Random Field theory or Monte Carlo simulation, for local maximum (or local minimum).

### 3.1 statistics for spatial association,

The two methods described above use two well known statistics describing associations: the chi-square of independence and the entropy, but are evaluated here on the contingencies of co-occurrences at a chosen order:

$$H_{Su}(C_{oo}, d) = -1/\log(N_{c_{oo}}) \sum_{c_{oo}}^{N_{c_{oo}}} p_{c_{oo}} \log(p_{c_{oo}}) \quad (1)$$

which defines a spatial entropy as the entropy, normalised to uniformity, of the multinomial distribution of co-occurrences at distance ***d***, with multi-index $c_{oo}$ according to indexes of the categorical variables (attributes) and the co-occurrence order;

$$\frac{\chi^2}{N} = \sum_{ijk} p_{i..}p_{.j.}p_{..k}\left(\frac{p_{ijk} - p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}}\right)^2 \quad (2)$$

which is the chi-square of a co-occurrence table of order 3 for 3 variables indexed by ***i***, ***j***, and ***k.***. Different ways of computing co-occurrences were described in Leibovici et al.(2008) but the simplest to understand is that the maximum distance between the co-occurrent points (with defined labels) is at most ***d***.

**3.2 scan of co-occurrences,**

Different strategies of scan can be suggested. For example, for each point $x_s$ of interest, a neighbourhood $Vx_s$ is built to reach a condition of sufficiency (*e.g.* the number of points in $Vx_s$ is exactly $n_1$, or the number of "cases" in $Vx_s$ is $n_2$), then the statistics above (one or the other) are computed within the neighbourhood and attributed to the point $x_s$. The condition of sufficiency is fundamental here in order to ensure comparisons of the values obtained. Notice that ***d*** (distance of collocation) also plays an important role here; adaptation or optimisation with the parameters of the condition of sufficiency is an important aspect of the method (*e.g.* a range of ***d*** can be explored and part of the result is the maximum ***d*** at which the chosen statistic reaches a significance threshold). In a similar approach to GAM, one could also set the size and shape of neighbourhood, or as with a spatial scan, the neighbourhood could grow until the statistic is maximised (or minimised, in the case of spatial entropy).

**4. Discussions**

This methodology can be implemented within a parallel computing environment and distributed network using web services compliant to OGC standards. Specifically, the point data may be provided through Web Feature Services and the multivariate computation provided through Web Processing Services. The multivariate approach of ScankOO can allow *mashups* of several datasets, though care must be taken about sample size and consistency of sampling among data sources. There is a trade off between ***k,*** the order of co-occurrence, which mainly acts like a spatial constrainer, and the necessary sample size to build the multiway contingency table which estimates the multiway-multinomial distribution. Depending on the health or epidemiological study, the scan strategy - and particularly the condition of sufficiency for the neighbourhood - may be quite different. Example implementations onreal data from an MRSA study already used in Leibovici et al.(2008) will be presented at the conference.

**References**

Besag J and Newell J (1991) The Detection of Clusters in Rare Diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **154**, 143-155.

Bivand R.S Pebesma E.J and Gómez-Rubio V (2008) *Applied Spatial Data Analysis with R*.1st Edition., Springer-Verlag, New York Inc.

Ceyhan, E (2009) On the use of nearest neighbor contingency tables for testing spatial segregation. *Environmental and Ecological Statistics.* OnlineFirst

Kulldorff M and Nagarwalla N (1995) Spatial disease clusters: Detection and inference. *Statistics in Medicine*, **14(8)**, 799-81

Kulldorff M. (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods*,

**26**:1481-1496.

Kulldorff M Tango T and Park P.J (2003) Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, **42(4)**, 665-684.

Kulldorff M (2006) Tests of spatial randomness adjusted for an inhomogeneity: a general framework**.** *Journal of the American Statistical Association* 2006, **101(475),** 1289-305

Lawson A Gangnon R and Wartenberg D (2006) Developments in disease cluster detection. *Special Issue: Statistics in Medicine* **25, (5)**

Leibovici D.G Bastin L and Jackson M (2008) Discovering Spatially Multiway Collocations. *GISRUK Conference 2008, Manchester, UK, 2-4 April, 2008*

Leibovici D.G (2009) Spatio-temporal Multiway Decomposition using Principal Tensor Analysis on k-modes: the R package PTAk. *Journal of Statistical Software* (accepted August 2009)

Leibovici D.G Bastin L and Jackson M (2009) Higher Order Cooccurrences in Point Pattern Analysis and Decision Tree Clustering. *Computers & Geosciences,* (submitted)

Openshaw S Charlton M Wymer C and Craft A.W (1987) A mark I Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, **1**, 335-358

Waller L.A and Gotway C.A (2004) *Applied Spatial Statistics for Public Health Data*. Wiley, Hobooken, NJ.

**Biography**

*Dr Didier Leibovici is a Research Fellow in geospatial modelling and analysis, with previous posts as a statistician in epidemiological/medical imaging research and as a geomatician for landscape changes in agro- ecology. Interests refer to interoperability and conflation models for cross-scales of integrated modelling applications within an interoperable framework chaining web services.*

*Dr Lucy Bastin is a Lecturer in GIS at Aston University. After a PhD on urban plant metapopulations, and research into fuzzy classification / uncertainty visualisation at Leicester University, she spent 3 years as a GIS software developer. Her current research interests include Web Processing Services for automatic interpolation, and spatial epidemiology.*

*Dr Suchith Anand is an Ordnance Survey Research Fellow at the Centre for Geospatial Science, University of Nottingham. His research interests are in open source GIS, automated map generalization, geohydroinformatics, mobileGIS, location based services, optimization techniques and asset management.*

*Dr Jerry Swan is a Research Fellow in Geocomputation. His approach includes optimization, graph theory, symbolic computation, knowledge representation, machine learning, semantics and ontologies. Jerry is particularly involved in the Persitent Test Bed project commissioned by OGC, AGILE and EUROSDR.*

*Dr Gobe Hobona is a Research Fellow in data, geoprocessing, workflow modelling standards and is consultant for the Open Geospatial Consortium (OGC) on the GIGAS project. Gobe's interests focus on management systems for OGC web services based on Grid computing and on cloud computing.*

*Pr. Mike Jackson is Director of the Centre for Geospatial Science. Prior to this he worked in industry at QinetiQ, Hutchison 3G, Laser Scan in various geospatial specialist and executive roles and in research for NERC. Mike is non-executive director of the Open Geospatial Consortium, and has research interests in combining new technologies such as positioning, pervasive computing and location based services for geo-informatics applications*

# Modelling health-harming behaviours in a socially ranked geographic space

Catherine Emma Jones[12],

[1]University College London, Gower Street, London, WC1E 6BT
[2]Department of Geography, University of Portsmouth,
Email: Kate-emma.jones@ucl.ac.uk

KEYWORDS: Health-harming Behaviours, Health and Lifestyle Status, Geodemographics, Health Geography

## 1. Introduction

The policy priority in the National Health Service (NHS) operating framework for 2009/10 calls for Primary Care Trusts (PCTs) to commission services based on the identification of local health outcomes connected to health-harming behaviours. To enable detection of local health needs they must first be successfully measured for different population groups. Indeed, measurement is fundamental to evidence-based policy, strategy and service delivery for both local and national scales but measuring health outcomes of behaviours is problematic. Current measurement is restricted to the use of administrative geographies, associated with the formal hierarchical scales of government and public sector organisations (Moon 1990). These formal boundaries produce a workable organisational structure practical for controlling budgets. Despite this, their usefulness for understanding the health needs of different groups of people is limited, due to their susceptibility to the ecological fallacy and their inherent arbitrary nature.

The determinants of health outcomes and the social environment of population groups are inextricably linked to geographical settings (Gatrell et al., 2004). This paper explores the practicalities of departing from traditional conceptualisations and aligning the scale of health measurement within the framework of functional space. Geographical inquiry at this scale will delineate breakpoints between spheres of influences in adjacent facilities or features (Longley et al., 2005, page 135). The aim is to develop a framework to enable the exploration of health behaviours as defined by health status, within the context of social and lifestyle status of neighbourhoods. This will help to maximise within neighbourhood similarity and ensure between neighbourhood differentiation is also maximised.

As cities grow and become more crowded, the diverse nature of the population emerges as an observable social pattern. This diversity has repercussions for health intervention planning, since health outcomes are not distributed equally. Understanding the nature of the diversity within the realm of functional scales will promote the identification of neighbourhoods where social groups are sharply differentiated in health and lifestyle status. This paper considers how geodemographic classifications, whilst not without limitations, can provide a geographic and analytical framework to investigate the social-spatial differentiations of health status that emerge when using functional geographies.

This paper explores the geographical and social space of functional neighbourhoods in reference to health-harming behaviours by assessing lifestyle and health status of the residential population. It appraises the usefulness of a spatial framework based on geodemographic classifications to facilitate the examination of health-harming behaviours. Using Bourdieu's perception of social space as a

starting point (Bourdieu, 1987), social space is explored using a measure of lifestyle status as a factor of the three different capitals; economic, cultural and social. It considers how health behaviour can be predicted by means of a composite measure of health status.

## 2. Methodology

Typically geodemographic classifications cluster small areas together, combining similar socio-economic variables to classify and order neighbourhoods into different social groups. One advantage these classifications have over traditional data is how they engage our understanding of demographic data and information according to social space. Recently, applications of geodemographic classifications have become more popular but their primary usage still remains in the private sector. Although they are increasingly being utilised in the public sector for education and crime, as yet little use has been made of them to differentiate population conditions of unhealthy lifestyles and behaviours.

To define health status, a commercial geodemographic classification (Mosaic: Experian, Nottingham, UK) was adjoined to the National Hospital Episodes database. Each patient episode was categorised by diagnosis defined by the International Classification of Diseases (ICD). Neighbourhood quotients were then calculated by averaging the total number of episodes[1] for a neighbourhood sub-group compared to the average number of total episodes for that sub-group in England (Webber, 2004). This enabled many types of health risks of being admitted to hospital to be investigated, even though not all illnesses can be attributed to lifestyle or behavioural choices. Those unrelated to lifestyle choices were identified and excluded from the analytical framework as they would be unsuitable for geodemographic analysis. To identify lifestyle related illness an index was developed based upon the extent of variation, for each individual illness, across different neighbourhoods. It is known as Total Weighted Deviation (TWD) and was first proposed by Webber (2004). It evaluates the suitability of population sub-groups to predict variation in illnesses, see equation 1.

$D_s$= Disease/condition index score

p = percentage of population

n = neighbourhood Type

$$TWD = \sum^{n=21}\left(\sqrt{(100-D_s)^2}\right)pn \qquad (1)$$

The results, Figure 1a, revealed the extent to which neighbourhoods deviate from that of the average neighbourhood. Illnesses with the highest TWD score represent differentiation within neighbourhoods associated with illnesses influenced by behaviour choices. The 10 illnesses with the most differentiation were then used as the input for the composite health status indictor. It was created by averaging the risk for each of the conditions in figure 1(a). The final measure predicted the likelihood of an individual in a neighbourhood to be exposed to health-harming illnesses, figure 1(b).

[1] Hospital Episode Statistics: A database of patient-based records of finished consultant episodes by diagnosis, operation and speciality from NHS hospitals in England (ONS, 2010)

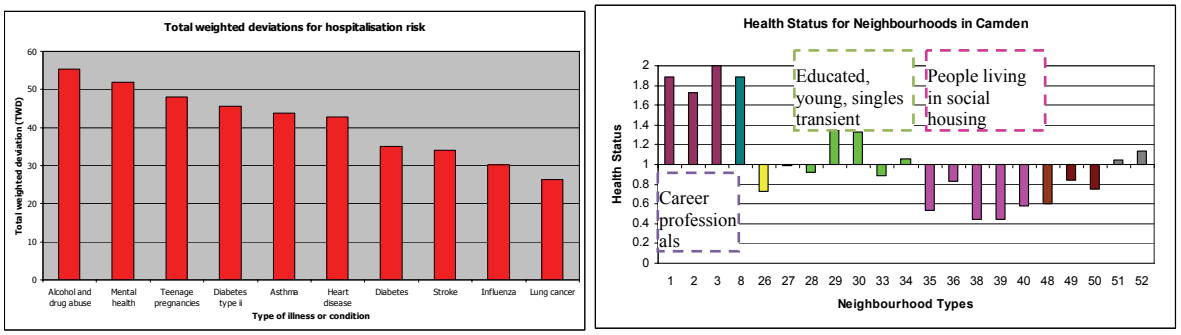



**Figure 1: (a)** Total weighted deviation scores for hospital risk of 10 illnesses and **(b)** Health status for neighbourhoods (note: health status diverges from 1; a score of 2 represents neighbourhoods with the healthiest lifestyle behaviours whereas 0 corresponds to the unhealthy behaviours.)

The next step was to define lifestyle status. Bourdieu recognized that a number of constituents interlock together to create social space (1997), comprising different capitals: economic, cultural and social, which fluctuate in volume and composition over time. It is these differences that distinguish population groups. Using this viewpoint and combining it with information derived from a geodemographic base, social space can be explored. Lifestyle status was defined using extracted variables from the geodemographic classification. These proxy measures were likened to the different conceptions of Bourdieu's capital: economic capital was measured using unemployment, salary and housing tenure, cultural capital with education attainment and the neighbourhood classes represented the different types of social capital. The domain of cultural capital based on the premise of cultural capital being knowledge and skills acquired by socialisation and education (Painter 2000)

## 3. Results

The use of socially similar neighbourhoods as the functional unit of analysis helps our understanding of the local social context and aids the exploration of health status. In Figure 2 a visualisation of the local population of one London borough, Camden, summarises the population health status alongside their lifestyle status. The pattern that emerges is familiar as it replicates the social gradient of health inequalities first proposed by Marmot and Wilkinson (2004). This pattern reinforces the notion that societal differences in health and lifestyle status persist, albeit to different extents, through the social spectrum.

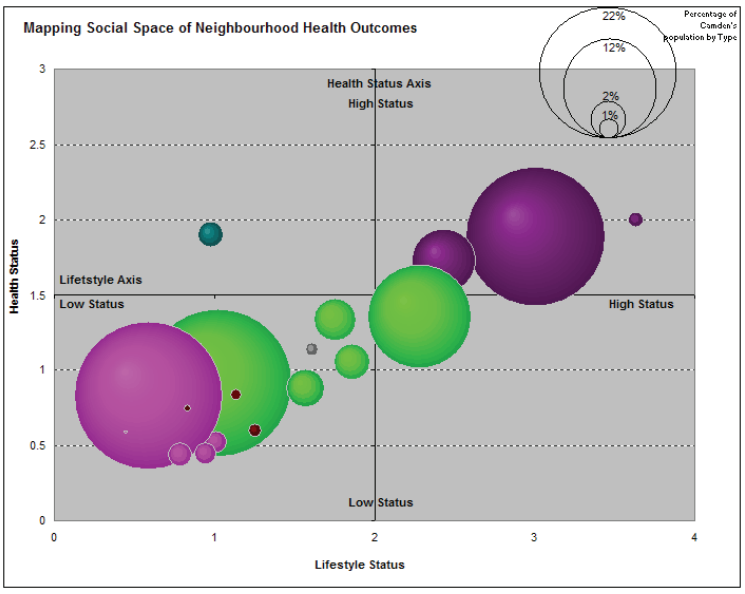
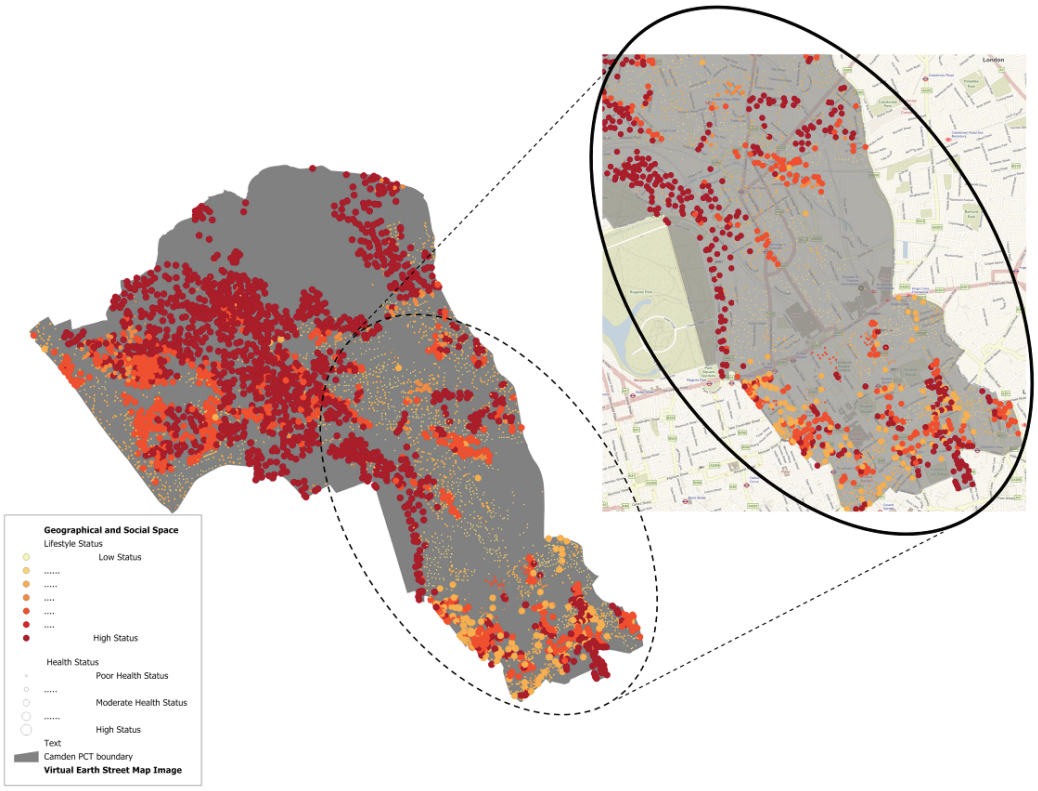


**Figure 2:** Health status of neighbourhoods in Camden, London, UK, plotted in social space.

The graph can be transformed into its geographical space to reveal spatial patterns. Figure 3 represents social similarity and geographic proximity for health/lifestyle status across different neighbourhoods. The size of the circle maps health status whilst its colour models lifestyle status. The map captures the presence of health and lifestyle




inequalities.

**Figure 3:** Representations of lifestyle and health status for Camden's neighbourhoods

(Note: Large red circles are indicative of a positive healthy lifestyle whereas small yellow circles indicate the most unhealthy health and lifestyle status)

Large disparities of lifestyle status and health inequalities are observed. Neighbourhoods with similar scores for lifestyle status do not have the same scores for health status, and conversely neighbourhoods with a shared social space do not always have the same predicted health status. Furthermore people in adjacent neighbourhoods occupy dissimilar social spaces and have diverging health status. This all points to an apparent contradiction to Tobler's First Law of Geography, where near objects are more related than far objects.

## 4. Discussion & Conclusions:

The different capitals modelled in this paper are cognisant of social determinants of health (Dahlgren and Whitehead, 1991) which include factors such as personal behaviour, lifestyles, community influences, living and working conditions and educational attainment. Measurement of these elements at the functional scale of health geographies aids recognition of behavioural patterns within our lived space. For health care policy and practice, intricate knowledge of both the geographical and social space, where and how we inhabit space, will aid clearer understanding of population groups, a fundamental requirement to effective health targeting and intervention planning.

In the literature there are concerns relating to the limitations of geodemographic typologies, which have not yet been discussed in this paper but briefly include over generalisation and reductionist (Goss, 1995; Curry, 1997), uncertainty (the levels of which are difficult to ascertain), subjection to the ecological fallacy (Weeks, 2004) as well as the misunderstanding that they are true depictions of reality and symbolise discrete distinct groups (they actually indicate social averages). These limitations withstanding, the location effects that arise from the geographical proximity of socially similar populations are evident in geodemographic clustering (Longley and Webber, 2003) and for the purposes of health care planning which includes social marketing they offer one method of enriching data at the local scale.

The work presented in this paper builds on Gatrell's ideas of social and geographical space and Bourdieu's belief that the diverse nature of social space requires distributions to account for "socially ranked geographical space" (Bourdieu, 1986, pg 124). Geodemographics were used to define this ranked view of the geographic and social space and to provide the underlying analytical framework. The neighbourhood, a sociological construct, was used as the representative unit of the functional scale within which people live. The use of geodemographics in this context provides the mechanism for maximising homogeneity of social processes within the different neighbourhoods. This work supports Moon (1990) and his desire for functional geography to be the basis for the community, neighbourhood or locality. For these reasons geodemographic classifications offer an alternative spatial framework for neighbourhood representations of social space.

Furthermore the 3 components of social space proposed by Sorre (1957) are shape, localisation and division. In this paper a geodemographic framework did enable observation of these components since to some extent they allude to the homogenous characteristics and social groupings of people with the potential to experience the same life and health status. The resulting measures in the form of averages indicated of predictive risk of being admitted to hospital. This provided both the health and social context to be embedded into this geodemographic framework. At this preliminary stage the measures of lifestyle and health status are somewhat crude, and are subject to the well document limitations that surround the use of neighbourhood typologies as the scale for analysis. Further work indicates the need to incorporate more complex measure of cultural diversity to provide further enrichment, The initial findings suggest it would be beneficial to conduct further work on their development – which should include a element of qualitative ground truthing of the results and the measurement and incorporation of the different social networks to into the framework, as it is one of the elements of social space that is currently missing. With all things considered the initial analysis is promising and indicates a potential alternative method for viewing the social-urban-health landscape of complex local populations.

## 5. Acknowledgements

The research outlined in this paper initially began as part of a Knowledge Transfer Partnership between University College London and Camden Primary Care Trust (PCT) as part of KTP programme number 37. I would like to thank my former PhD Supervisors Dr Muki Haklay and Prof Paul Longley for their input and support. With thanks also to Experian and Richard Webber for proving some of the data for this study.

## 6. References and Citations

BOURDIEU, P., 1986. *The forms of capital*, New York: Greenwood Press.

BOURDIEU, P., 1997. The forms of capital. In Halsey, A.H., Lauder, H., Brown, P., and Wells, A.S., (eds), *Education: Culture, Economy, Society*, Oxford: Oxford University Press.

CURRY, M., 1997. The digital individual and the private realm. *Annals of the Association of American Geographers,* 87, pages 681 to 699.

DAHLGREN, G., AND WHITEHEAD, M., 1991. *Policies and strategies to promote social equity in health*, Stockholm: Institute of Future Studies.

GATRELL A.C., POPAY, J., AND THOMAS, C., 2004. Mapping the determinants of health inequalities in social space: can Bourdieu help us? *Health and Place*, 10, 245 - 257.

GOSS, J., 1995. We know who you are and we know where you live: the instrumental rationality of geodemographic systems. *Economic Geography*, 71, 171 - 198.

LONGLEY, P.A., 2005. Geographical Information Systems: a renaissance of geodemographics for public service delivery, *Progress in Human Geography*, 29,1 57 - 63

LONGLEY,P.A., WEBBER,R. (2003). Geodemographic analysis of similarity and proximity: their roles in the understanding of the geography of need. in Longley,P.A., Batty,M. (ed.) *Advanced Spatial Analysis*. Redlands, California: ESRI Press, 233-266

MOON, G., 1990. Conceptions of space and community in British health policy, Social Science & Medicine, 30, 1, 165 - 171.

PAINTER J. 2000. Pierre Bourdieu. In Thinking Space, ed. M Crang, N Thrift, pp. 239-59. London: Routledge

SORRE. 1957. *Rencontres de la géographie et de la sociologie*. Paris Librairie: Marcel Rivière.

WEBBER, R., 2004. *Neighbourhood inequalities in the patterns of Hospital Admissions and their application to the targeting of health promotion campaigns.* UCL: CASA Working Paper 90. [Online] http://eprints.ucl.ac.uk/1291/ [accessed 27th April 2008]

WEBBER, R., 2004. *The relative power of geodemographics vis a vis person and household level demographic variables as discriminators of consumer behaviour*, UCL: CASA Working Paper 84. [Online] http://www.casa.ucl.ac.uk/publications/workingPaperDetail.asp?ID=84 [accessed 26th December 2007]

WEEKS, J. R. (2004) The Role of Spatial Analysis in Demographic Research. In Goodchild, M.F., and Janelle, D.G., (eds), *CSISS best practice publications: spatially integrated social science*, Oxford: Oxford University Press.

## Biography

*Kate (Catherine) Jones completed her PhD in Geographies of Health in 2008. Having spent 2 years as a Post Doc at UCL she now has a lectureship in Human Geography at the University of Portsmouth. In the domain of health she is keen to peruse research into how the access, quality and visualisation of spatial data can be used for health intervention planning.*

# Informing Population Genetics through Spatial Analysis of Surnames

James Cheshire[1], Pablo Mateos[1], Paul A. Longley[1]

[1]Department of Geography and Centre for Advanced Spatial Analysis, University College London.

james.cheshire@ucl.ac.uk, spatialanalysis.co.uk

KEYWORDS: Surnames, Britain, Genetics, Location Quotients, Kernel Density Estimation.

## 1. Introduction

There is an established, and growing, literature demonstrating the links between geography, surnames and patrilineal (inherited from the father) genetic structures of populations (Jobling and King, 2009). When combined with historical population registers these create an opportunity to inform large-scale genetic sampling. This abstract will outline the initial stages of integrating spatial analysis with sampling from the People of the British Isles (POBI) project (http://www.peopleofthebritishisles.org/).

The POBI project is seeking to establish historical patterns of population movement and to determine the influence of other countries on the genetic makeup of people in contemporary Britain. To achieve these aims, genetic samples need to be taken from volunteers with a long family lineage in a particular area (Manni et al., 2005). Surnames provide useful criteria on which to establish if this is the case.

For many surnames in Great Britain their areas of origin remain the areas of highest concentration (Mascie-Taylor and Lasker, 1981). A discernable core locality and the link to a person's genes provides a powerful combination in genetic sample design (Bowden et al., 2008). Surnames outside of their area of origin, as defined by a core locality of highest concentration, are likely to be possessed by people whose ancestors originated elsewhere. The POBI participants with non-local surnames can therefore be removed from the sample as their genetic make-up may have been subject to different influences.

The two methods used- kernel density estimation (KDE) and location quotients (LQ)- to determine a surnames' core locality in this study were selected for their intuitiveness to those unfamiliar with spatial data and their ease of implementation using the R environment for statistical computing and graphics (R Core Team 2009). The widespread use and free availability of R was an important consideration as it is rarely the case that access (both in terms of provision and expertise) to commercial GIS platforms extends into non-GISc disciplines. The innovation of this study is not necessarily the methods, but their application to novel data and as part of an inter-disciplinary study.

## 2. Data

Although it is possible to determine a surname's area of origin from contemporary data, historical datasets are advantageous because they are less affected by recent migrations. The digitisation of the 1881 Census of Great Britain (UK Data Archive, www.data-archive.ac.uk/) therefore provides an invaluable resource in this context. The census provides the names and place of enumeration (Parish and Registration District) for 29 million people, with a total of 425,000 unique surnames (approximately 49,000 of which have occurrences of more than 20 people). The census has been geocoded to Registration District (RD) level (mean population 4900) and joined to a shapefile containing the boundary data (http://edina.ac.uk/ukborders/).

Proximity to a surname's core locality was determined for each of the places of birth of 842 POBI participants (with a total of 646 unique surnames). Participants should have been born in a rural area and within 60km of their maternal and paternal grandparents' places of birth to be eligible to participate in the research. This criterion was met by 625 out of the 842 participants (a further 196 had 3 out of 4 of their grandparents within the 60km threshold). Mean coordinates derived from the birthplaces of each participant and their grandparents within the 60km threshold were used in the analysis. The distribution of these points is shown in Figure 1.

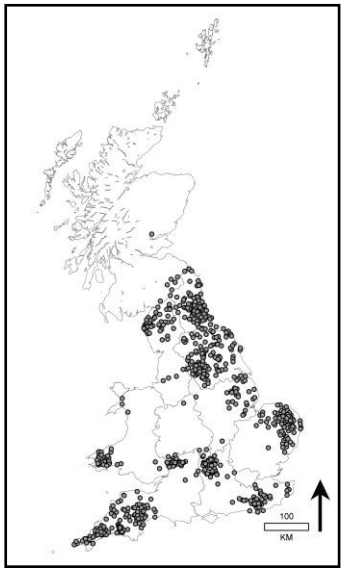



**Figure 1:** The geographical localities of the 842 participants in the POBI project. These locations are derived from the mean coordinates of the participant's place of birth and the places of birth of their 4 grandparents. Boundary data Crown Copyright Ordnance Survey.

### 2.1 Surface Analysis: Kernel Density Estimation

Kernel density estimation (KDE) was implemented to calculate the density of occurrences for the surname of interest in a neighbourhood around each occurrence in turn. KDE overlays points representing each occurrence (RD centroid) onto a grid before fitting a symmetrical smoothly curved surface (known as the kernel) over each of them (de Smith et al., 2007). The grid cells are assigned a value from the point on the kernel directly above their centroids. Each point is assigned its unique kernel; where kernels overlap from neighbouring points their sum is assigned to the corresponding grid cells. Cells near to, or containing, a large volume of points will therefore be ascribed higher values (de Smith et al. 2007). A fixed normally distributed kernel was used for each name. Each surname will have a varying extent so the size of the kernel cannot be fixed across all names. It's extent is therefore selected by normal optimal smoothing. See Bowman and Azzalini (1997, P.31) for more details. Areas with a high population density are likely to have higher frequencies of a particular name regardless of their geographic location. To account for this each of the surname points are weighted by the surname's relative frequency within the RD. The final stage in the methodology partitioned the density surface along a contour line at a specified interval to create the surname "core". This is a simple but effective means of adjusting the sensitivity of the analysis. For example if a very small locality is required then a contour at a high threshold value can be specified.

### 2.2 Discrete Analysis: Location Quotients

The location quotient (LQ) is an index that compares a region's share of a particular activity, in this case surname occurrence, with the share of that same activity found at a more aggregate spatial level (Burt et al., 2009). It can be defined as follows:

$$LQ_i^j = \frac{A_i^j / \sum_{i=1}^{n} A_i^j}{B_i^j / \sum_{i=1}^{n} B_i^j} \quad (1)$$

where $A_i^j$ is the frequency of surname *i* in region *j*, $B_i^j$ is the frequency of surname *i* in Great Britain and *n* represents the total number of surnames. LQ values of greater that 1 represent an RD with a higher concentration of the selected name than would be expected if the surname had a uniform distribution across Britain.

The RDs with the three highest location quotients are taken to be the surname's core locality. In many cases these are spatially contiguous or within close proximity. If this is not the case the surname may have multiple cores throughout Britain or an extremely dispersed distribution.

## 3. Results and Discussion

Table 1 gives three examples from the 842 output rows produced from the KDE analysis. KDE's primary limitation is evident from the number of districts included within the surname core. The circular nature of the kernels placed over each surname occurrence tends to include districts that do not contain that surname. In the case of some rare surnames the number of core RDs exceeds the total population of that name. The extent of this limitation was not anticipated and created a problem for the determination of realistic localities for many of the surnames in the sample.

**Table 1:** The Metrics Produced by Using KDE to Establish a Surname's Core Locality.

| Sample_ID | Name | Error_Status | No_of_Cores | Mean_Dist_Btwn_Cores (KM) | Inside_Core? |
|---|---|---|---|---|---|
| BAN010 | TAYLOR | OK | 1 | NA | 0 |
| NEA016 | TULLY | OK | 2 | 98.58 | 1 |
| CUM017 | BERTRAM | OK | 1 | NA | 0 |
| **Min_Dist_to_Core(KM)** | **Core_Area (KM)** | **No._Core_Districts** | **1881_Name_pop** | **Within150km?** | **Reason_For_Error** |
| 561.76 | 86440.71 | 415 | 191101 | OK | NA |
| 36.03 | 91739.46 | 42 | 1620 | OK | NA |
| 121.88 | 40902.06 | 76 | 1359 | !Check Map! | NA |

The LQ outputs are much better representations of a surname's spatial distribution. Table 2 provides two contrasting outputs from the LQ methodology. COR001 would be classified as a sample taken close to or within the Dunstan surname's locality. It has a high core LQ (140.07), relatively few significant districts (those with an LQ >1), a short distance (9.6km) to the centroid of the core RD, and a low mean distance between the three districts with the highest LQs (11.12km) suggesting a spatially clustered distribution of the districts with the highest concentration of that name. COR002 would be classified as a sample without a core surname locality. The surname has a large total population, relatively low LQs for the top three districts, and a large distance between these districts.

**Table 2:** The Metrics Produced by Using LQs to Establish a Surname's Core Locality.

| Sample_ID | Surname | Sample_X_Coord | Sample_Y_Coord | GB_Surname_Pop |
|---|---|---|---|---|
| COR001 | DUNSTAN | | | 1694 |
| COR002 | RICHARDS | | | 45120 |
| **Core_District** | **Core_X_Coord** | **Core_Y_Coord** | **Name_Pop_Core** | **Tot_Core_Pop** |
| Falmouth | 176509.48 | 32561.68 | 201 | 25103 |
| Helston | 168815.84 | 25695.16 | 699 | 23691 |
| **Core_LQ** | **N_Sig_Districts** | **Dist_To_Core (KM)** | **Highest_LQ_GOR** | **Core_District2** |
| 140.07 | 71 | 9.63 | South West | Redruth |
| 19.38 | 166 | 39.70 | Wales | Aberystwith |
| **Core_X2_Coord** | **Core_Y2_Coord** | **Name_Pop_Core2** | **Tot_Core_Pop2** | **Core_LQ2** |
| 166396.44 | 40334.36 | 351 | 46106 | 133.17 |
| 267995.17 | 280490.23 | 603 | 25526 | 15.51 |
| **Dist_To_Core2** | **Core_District3** | **Core_X3_Coord** | **Core_Y3_coord** | **Name_Pop_Core3** |
| 5.93 | Helston | 168815.84 | 25695.16127 | 127 |
| 236.69 | Redruth | 166396.44 | 40334.35919 | 1013 |
| **Tot_Core_Pop3** | **Core_LQ3** | **Dist_To_Core3** | **Mean_Dist_Btwn_Cores (KM)** | **Comment** |
| 23691 | 93.78 | 8.92 | 11.13 | OK |
| 46106 | 14.43 | 32.55 | 101.03 | OK |

The highest LQs for the majority of surnames are less than 200 and their three core RDs are less than 200km from each other. This suggests that while few core districts are likely to be contiguous, the majority of surnames can be classified as having a single core region. This is validated by the KDE results that suggested around 9% of the surnames in the sample have multiple cores. Figure 2 provides a mapped illustration of two outputs from the analysis.

Unlike the KDE, LQs enable the RDs to be ranked according to the relative concentration of the surname. LQs also exclude districts that are close to the surname core but do not contain that name. This overcomes some of the KDE problems outlined above. The drawback to this is of course the discrete handling of what can often be considered a continuous phenomenon (Lasker and Mascie-Taylor, 1985). A key strength of both methodologies, however, is their intuitiveness and relative simplicity. Many of the results will be communicated and analysed by researchers unfamiliar with spatial data and would therefore be loose meaning if they were conceptually hard to understand and interpret.

Finally, the metrics produced can only be as good as the data they are derived from. In terms of the primary data the process of geocoding samples was inconsistent due to missing postcodes, and specific addresses- especially for the ancestral places of birth. These localities were therefore geocoded at range of scales- from the centroids of a postcode to, in rare cases, the centroids of counties. A further limitation is the shortcomings of the 1881 census. Although an invaluable resource, some surnames are likely to be over-represented and others (most likely through enumerator and subsequent digitisation errors) under-represented. Whilst the process of selecting the final genetic samples remains largely subjective, the metrics produced above provide a richer source of information than has previously been available in genetic studies of Great Britain.

This preliminary research has sought to address the need for spatial analysis in genetic sampling design. It has shown the utility of both surnames and simple spatial analysis in this context through the calculation of kernel density estimates and location quotients. Much work remains to be done in this area that can provide fruitful collaborations between GISc and genetics.

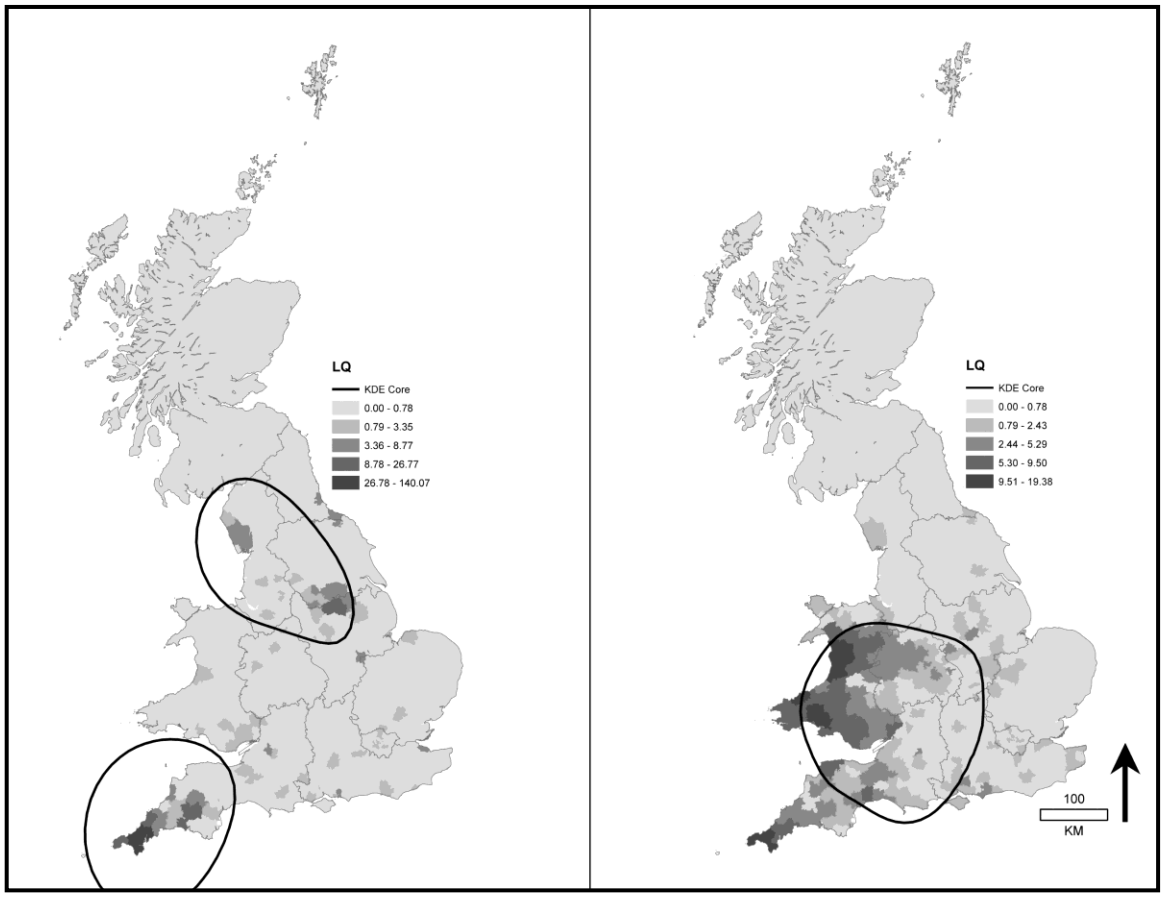



**Figure 2:** Maps demonstrating the core localities derived from the LQ and KDE methodologies for the surnames Dunstan (left) and Richards (Right). Boundary data Crown Copyright Ordnance Survey.

## 4. Acknowledgements

The authors would like to thank Bruce Winney and Kristin Nicodemus from the POBI project for the provision of data and feedback on the methodologies employed above. This project was undertaken as part of James Cheshire's ESRC CASE PhD Studentship in Collaboration with ESRI (UK). The research contributes to a larger Wellcome Trust funded project in collaboration with Oxford University. The reviewer's comments were gratefully received and we hope have been fully addressed in the improvements made to the original paper and its associated presentation.

## 5. References

Bowden, G., Balaresque, P., King., T. et al. 2008. Excavating Past Population Structures by Surname-Based Sampling: The Genetic Legacy of Vikings in Northwest England. *Molecular Biology and Evolution.* 25, 2: 301-309.

Bowman, A., Azzalini, A., 1997. *Applied Smoothing techniques for Data Analysis.* Clarendon Press, Oxford.

Burt, J., Barber, G., Rigby, D. 2009. *Elementary Statistics for Geographers* . Third Edition. The Guilford Press, London.

de Smith, M., Goodchild, M., Longley, P. 2007. *Geospatial Analysis*. Second Edition. Matador, Leicester. Available online: www.spatialanalysisonline.com.

Jobling, M, King, T. 2009. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends in Genetics.* 25, 8: 351-360.

Lasker, G., Mascie-Taylor, C., 1985. The Geographical Distribution of Selected Surnames in Britain. Model Gene Frequency Clines. *Journal of Human Evolution,* 14: 385-292.

Manni, F., Toupance, B., Sabbagh, A., Heyer, E. 2005. New Method for Surname Studies of Ancient Patrilineal Population Structures, and Possible Application to Improvement of Y-Chromosome Sampling. *American Journal of Physical Anthropology.* 126: 214-228.

Mascie-Taylor, C., Lasker, G. 1984. Geographical Distribution of Common Surnames in England and Wales. *Annals of Human Biology*. 12, 5: 397-401.

## 6. Biographies

***James Cheshire*** *is halfway through his ESRC CASE PhD studentship (in collaboration with ESRI (UK)) in UCL's Department of Geography. His research focus is the spatial analysis of surnames and its applications. He is a also a research assistant on the Wellcome Trust's "People of the British Isles" project. James' research can be followed at spatialanalysis.co.uk.*

***Paul Longley*** *holds a chair in Geographic Information Science at UCL and acts as Deputy Director of CASA. His publications include twelve books and more than 125 refereed journal articles and contributions to edited collections. He is a co-I on a node of the ESRC National Centre for E Social Science and a co-editor of the journal Environment and Planning Series B.*

***Pablo Mateos*** *is Lecturer in Human Geography in the Department of Geography at University College London (UCL). His research interests lie within Population and Urban Geography and his work focuses on investigating ethnicity, migration and socio-spatial inequalities in contemporary cities. PhD in Geography (UCL 2007); MSc in GIS and Human Geography (University of Leicester 2004).*

# Applying network analysis to quantify urban greenspace accessibility for different socio-economic groups

Fariba Sotoudehnia[1], Alexis Comber [2]

1 University of Leicester, Department of geography, LE1 7RH;

Tel. 0116 252 3849,

Email: fs69@le.ac.uk; www.le.ac.uk/gg/staff/pg_sotoudehnia.html

2 University of Leicester, Department of geography, LE1 7RH; Tel. 0116 252 3854, Email: ajc36@le.ac.uk; www.le.ac.uk/gg/staff/academic_comber.html

KEYWORDS: Accessibility, Urban Greenspace, Deprivation, Social justice

## 1. Introduction

Provision of and access to urban greenspace is important for a number of reasons including: social interactions (Volker, 2006; Milton, 2002); physical health and psychological wellbeing (Maas et al., 2008; Grahn and Stigsdotter, 2003) especially for children (Francis, 2006), economical benefits (Jim and Chen, 2006; Mansfield et al., 2005) and environmental services (Yang et al., 2005; Fang and Ling, 2003).

The necessity of providing public with equitable accessibility to greenspace becomes particularly important under the notions of social and environmental justice due to the significant role of the areas in enhancing community identity, community attachment and improving local landscape (Matsuoka and Kaplan, 2008). Other work considering accessibility has also noted that, people who live in close proximity to greenspace have more chance to use areas frequently (Hoehner et al., 2005; Tyrväinen et al., 2004; Van Herzele and Wiedemann 2003) which contributes to peoples' health and quality of life (Pinder et al., 2009; Santos et al., 2009; Mitchell and Popham, 2008).

Geographical information systems (GIS) have been used to explore the social and environmental justice aspects in relation to spatial distribution of urban greenspace. For example, Kessel et al., (2009) used GIS to characterise access to greenspace in distance terms, and how such access has changed between 1990 and 2003 based on people socio-economic status. Work by Comber et al., (2008) used network distances to analyse greenspace accessibility for different ethnic and religious groups. Barbosa et al., (2007) measured accessibility to public greenspace to households in Sheffield, and examined how this varies across different sectors of society. Heynen et al., (2006) analysed the spatial distribution of urban greenspace against income. Omer and Or (2004) used "coverage approach" to examine the differential access to urban greenspace among Arab and Jewish population in two mixed Israeli cities. Neuvonen et al., (2007) studied the relationship between access to greenspace and the frequency of visits in Helsinki. Oh and Jeong (2007) used the network analysis method of GIS, to analyse pedestrian accessibility to urban parks in Seoul and the subsequent serviceability of the parks.

Reviewing literature reveals, although a good deal of attention in the current literature has been paid to social and environmental justice relating to spatial access to urban greenspace, more research is particularly needed to substantiate physical access against socio-economic status from both qualitative and quantitative aspects. To fill in the current gap, this study aims to use GIS-based network analysis method besides initial qualitative methods to answer how do physical access and perception towards urban greenspace vary across different socio-economic groups?

## 2. Methods

### 1.2. Study area

The city of Leicester with the total population of 280,000 located in the East Midlands in England is the study area to quantify accessibility against socio-economic status. The total area of the city (18,060 acres, 73.09 Km2), supporting almost 10% (2,000 acres) parks and open spaces, 6% (1100 acres) nature reserves and approximately 27% (4500 acres) as gardens.

### 2.2 Definition

In this study urban greenspace, accessibility and people socio-economic status are defined as bellow:

"*Urban greenspace*": publicly owned and publicly accessible open space in urban context where are covered by high degree of vegetation, e.g. urban parks, woodlands, spinney, meadows and other type of greenspace.

"*Accessibility*": physical distance or walking time from a residential home to an urban greenspace.

"*Socio-economic status*": is addressed according to the deprivation. Deprivation is a measure of poverty based on a number of criteria such as economic circumstances, health, crime, housing, educational achievement, skills and the environment. This study uses Townsend Index devised by Townsend et al. (1988) to quantify deprivation relating to public census data (i.e. unemployment, overcrowding, non car ownership, and non home ownership).

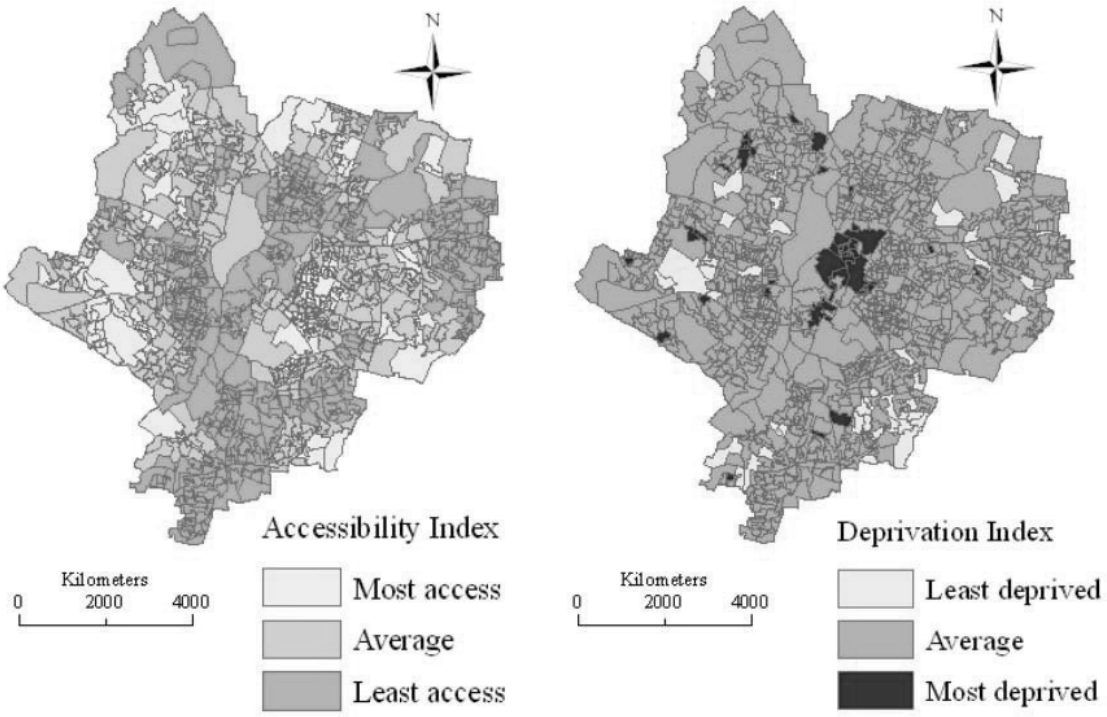
### 3.2 Data collection

The greenspace data of Leicester was provided by Leicester City Council. The output areas polygons (OAs) data and road network data, was extracted from UKBOURDERS and Ordnance Survey Meridian 2 (1:50,000), Edina. Census data for each OAs was obtained for the 2001 census from CASWEB service. Townsend Index was calculated according to unemployment, overcrowding, non car ownership, and non home ownership and the spatial distribution of deprivation was identified according to the 5 % threshold from the top and bottom levels of Townsend Index values.

### 4.2 Data analysis

Physical access was analysed in terms of shortest distance line from each residential output areas to the nearest greenspace access point in urban context. The underlying premise was that shorter distances are associated with greater and more frequently use of the areas (Hoehner et al., 2005; Tyrväinen et al., 2004; Van Herzele and Wiedemann 2003). Accordingly, the closest facility option in network analysis extension was applied to quantify different physical access points within 5 minutes (equal to the 300 metres), between 5 and 10 minutes (almost equal to the 1000 meters) and more than 15 minutes (distance more than 1000 meters).

With refer to the specified access criteria, Figure 1a shows the physical pattern of access to greenspace in Leicester categorising into "least access", "average" and "most access". The grey colour lines on the map show the boundary areas of each OAs. To show the spatial distribution of deprivation, the attribute table of Townsend Index values was joined to the attributes table of output areas polygons. Figure 1b shows distribution of deprivation in Leicester according to "5% least" and "5% most" values of Townsend Index. Rating colours from yellow to dark brown shows the characteristic of each OAs relating to the spatial deprivation as "least deprived", "average" and "most deprived".




a) b)

**Figure 1** The spatial distribution of a) access to urban greenspace in Leicester and b) the most (top 5%) and least (bottom 5%) deprived

### 3. Results

The initial results of overlaying accessibility against deprivation characterise Leicester's OAs into one of the nine following categories, based on the population (Table 1).

**Table 1** Provision of physical access against deprivation in Leicester (Sotoudehnia, 2010)

| Deprivation | Most access | Average | Least access | Total Grand |
|---|---|---|---|---|
| Least deprived | 2635 | 40821 | 988 | 44444 |
| Average | 6586 | 99855 | 4147 | 110588 |
| Most deprived | 5175 | 113103 | 6611 | 124889 |
| Total | 14396 | 253779 | 11746 | 279921 |

The Mosaic plots method – first proposed by Hartigan and Kleiner (1981) and extended later on by Friendly (1994) – was used to assess the relative equity of physical access to greenspace amongst different social groups. The deprivation scores were used to determine the set of "5% most deprived" and "5% least deprived". The numbers of people with and without access greenspace (i.e. <300m) in each census output area were summed for the different deprived groups. Figure 3 visualises the results of mosaic plot.

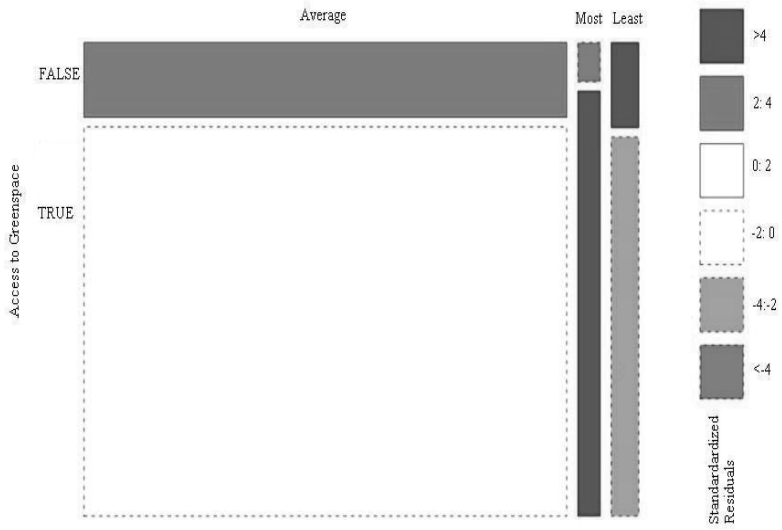



**Figure 1** Mosaic plot of access to urban greenspace by deprivation

The coloured plots show which social groups are under or over represented. The blue tiles show combinations of access and deprivation that are higher than average. The tiles shaded deep blue correspond combinations of access and deprivation whose residuals are greater than +4, when compared to a model of proportional equal levels of access for all deprived groups. This indicates a much greater frequency in those cells than would be found if this model were true. The tiles shaded deep red correspond to the residuals less than -4 indicating much lower frequencies than would be expected.

As a result, the mosaic plot method indicates that people, those placed into the most deprived group has significantly "*more access*" to greenspace than expected. Whilst the other derived groups (average and least deprived) have "*less access*" than would expected under an assumption of equal physical accessibility.

In order to estimate likelihood of access, "generalised linear models" were used as a function of deprivation. The results reveal that in the more deprived (5%) parts of Leicester, there is significantly "*more access*" to greenspace than 90% of the city and that the least deprived areas (5%) have "*less*" levels of physical access to greenspace.

## 4. Discussion and Conclusion

This study analysed physical access to urban greenspace against national access guidelines, which recommend no person should live further than 300m from a greenspace. "Network analysis" used as an equipped and powerful analytical tool to quantify physical accessibility. The Mosaic plot approach employed to visualise the provision of physical access to greenspace across different social groups.

In the light of the study's objective – quantifying accessibility against public socio-economic status relative to social and environmental justice – the results discolour the assumption of social justice in provision of and access to greenspace in Leicester. The results indicate the converse influence of deprivation on physical access by exploring that, in Leicester, people from the most deprived areas has significantly more accessibility to greenspace than expected and other derived groups (average and least deprived) have less access than would expected under an assumption of equal access. These results support the previous findings of Kessel, et al. (2009) indicated people from more deprived areas have better access to Thames Chase Community Forest (TCAF). In contrast to these findings,

Omer and Or (2005), addressed people from more affluent areas have greater access to greenspace in Israeli cities (Omer and Or 2005).

Although this study determined accessibility by physical proximity, parameters such as size and people's experience of attractiveness and appropriateness of greenspace are likely to be important whilst as yet has received relatively little attention (Balram and Dragicevic, 2005; Dwyer and Childs, 2004). This raises the necessity of conducting further work in this area to study the influence of deprivation on the way greenspace use and perceived by public in the city of Leicester. Such integrated analyses are particularly important to planning process and to support social and environmental justice objectives.

## 5. References

Balram, S., Dragicevic, S., (2005) Attitudes toward urban green spaces: integrating questionnaire survey and collaborative GIS techniques to improve attitude measurements *Landscape and Urban Planning* **71** pp 147-162

Barbosa, O. Tratalos, J. A., Armsworth, P.R., Davies, R.G., Fuller, R.A., Johnson, P., Gaston K.J., (2007) Who benefits from access to green space? A case study from Sheffield, UK *Landscape and Urban Planning* **83** pp 187-195

Comber, A., Brunsdon , C., Green, E., (2008) Using a GIS-based network analysis to determine urban greenspace accessibility for different ethnic and religious groups *Landscape and Urban Planning* **86** pp 103-114

Dwyer, J.F., Childs, G.M., (2004) Movement of people across landscape: A blurring distinction between areas, interests, and issues affecting natural resource management *Landscape and Urban Planning* **69** pp 153-164

Fang, C.F., Ling, D.L., (2003) Investigation of the noise reduction provided by tree belts *Landscape and Urban Planning* **63** pp 187-195

Francis. M., (2006) Urban parks as community places University of California pp1-10. Available at: www.lda.ucdavis.edu/people/websites/francis/korea.pdf. (Accessed 2008)

Friendly, M., ( 1994) Mosaic displays for multi-way contingency tables *Journal of the American Statistical Association* pp190-200

Grahn, P., Stigsdotter, U.A., (2003) Landscape planning and stress *Urban Forestry and Urban Greening* **2**, pp 1-18

Hartigan, J. A., Kleiner, Mosaics for contingency tables (1981) In: W. F. Eddy (Ed.) Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface New York: Springer-Verlag

Heynen, N., Perkins, H.A., Roy, P., (2006) The political ecology of uneven urban greenspace: the impact of political economy on race and ethnicity in producing environmental inequality in Milwaukee *Urban Affairs Review* **42(1)** pp 3-25

Hoehner, C.M., Brennan Ramirez, L.K., Elliott, M.B., Handy, S.L., Brownson, R.C., (2005) Perceived and objective environmental measures and physical activity among urban adults *American Journal of Preventive Medicine* **28** pp105-365

Jim, C.Y., Chen, W.Y., (2006) Impacts of urban environmental elements on residential housing prices in Guangzhou (China) *Landscape and Urban Planning* **78** pp 422-434

Kessel, A., Green, J., Pinder, R., Wilkinson, P., Grundy, C., Lachowycz, K (2009) Multidisciplinary research in public health: A case study of research on access to greenspace *Public Health* **123** pp 32-38

Maas, J., Verheij, R.A., Spreeuwenberg, P., Groenewegen, P.P., (2008) Physical activity as a possible mechanism behind the relationship between green space and health: A multilevel analysis *BMC Public Health* **8 (206)** doi:10.1186/ 1471-2458-8-206.

Mansfield, C., Pattanayak, S.K., McDow, W., McDonald, R., Halpin, P., (2005) Shades of green: Measuring the value of urban forests in the housing market *Journal forest economics* **11** pp 177-199

Matsuoka, R.H., Kaplan, R., (2008) People needs in the urban landscape: Analysis of Landscape and Urban Planning contributions *Landscape and Urban Planning* **84** pp 7-19

Milton, K., (2002) *Loving nature: Towards ecology of emotion* Rutledge, New York.

Mitchell, R., Popham, F., (2008) Effect of exposure to natural environment on health inequalities: An observational population study *The Lancet* **372 (9650)** pp1655-1660

Neuvonen, M., Sievanen, T., Tonnes, S., Koskela, T., (2007) Access to green areas and the frequency of visits – A case study in Helsinki *Urban Forestry and Urban Greening* **6** pp 235-247

Oh, K., Jeong, S., (2007) Assessing the spatial distribution of urban parks using GIS *Landscape and Urban Planning* **82** pp 25-32

Omer, I., OR, U., (2005) Distributive environmental justice in the city: Differential access in two mixed Israeli cities *The Royal Dutch Geographical Society* **96(4)** pp 433-443

Pinder R, Kessel A, Green J, Grundy C., (2009) Exploring perceptions of health and the environment: A qualitative study of Thames Chase Community Forest *Health and Place* **15(1)** pp 349-356

Santos, M.P., Page, A.S., Cooper, A.R., Ribeiro, J.C., Mota, J., (2009) Perceptions of the built environment in relation to physical activity in Portuguese adolescents *Health and Place* **15 (2)** pp 548-552

TOWNSEND, P., PHILLIMORE, P., BEATTIE, A. (1988) *Health and Deprivation: Inequality and the North*. London: Croom Helm.

Tyrväinen, L., Mäkinen, K., Schipperijn, J., and Silvennoinen, H., (2004) Mapping social values and meanings of green areas in Helsinki, Finland *Department of Forest Ecology, University of Helsinki, Finland.* Available at: www.sl.kvl.dk/euforic/nbw.htm (Accessed 2009)

Van Herzele, A., Wiedemann, T., (2003) A monitoring tool for the provision of accessible and attractive urban green spaces *Landscape and Urban planning* **63** pp 109-126

Volker, B., (2006) A walks in the park *GreenSpace*, Available at: www.green- space.org.uk (Accessed 2009)

Yang, J., McBride, J., Zhou, J., Sun, Z., (2005) The urban forest in Beijing and its role in air pollution reduction *Urban Forestry and Urban Greening* **3** pp 65-78

**Biography:**

**Fariba Sotoudehnia**

I am a Ph.D student in GIS and Human Geography at the University of Leicester. My current research concerns investigating greenspace accessibility through combining a GIS-based network analysis with different qualitative methods involving PPGIS, interviews and questionnaire. I am working under supervision of Dr Alexis Comber.

**Dr Alexis Comber**

Dr Alexis Comber is a senior lecturer in Geographic Information at the University of Leicester and his key research interests are:

- Analysis of uncertainty in geographic information and spatial data
- Methods for reasoning under uncertainty: Dempster-Shafer Theory of Evidence, Bayesian Probability, Endorsement Theory
- Land Cover, Land Use, Habitats
- Data Quality, Metadata, Data Standards
- Spatial analysis of policy and planning
- Accessibility, equity of access
- Optimisation, Genetic Algorithms

# An Exploration of Volunteered Geographic Information Stakeholders

Christopher J. Parker[1], Andrew May[2], Val Mitchell[3]

[1]Ergonomics and Safety Research Institute (ESRI)
Garendon Building, Loughborough University, Loughborough, LE11 3TU
+44 (0)1509 22 6900
Email: C.Parker@lboro.ac.uk, www.UserGeneratedDesign.co.uk
[2]A.J.May@lboro.ac.uk
[3]V.A.Mitchell@lboro.ac.uk

KEYWORDS: volunteered geographic information, neogeography, human factors, stakeholders, user-generated content.

**Abstract**

Volunteered Geographic Information (VGI) has huge potential for influencing the use of geographic information systems. However, there is a wide range of individuals involved in this process, each with their own motivations for contributing and using volunteered data. This paper investigates the range of stakeholders involved with VGI, their relationships and the main tensions and issues involved. The research was based on a series of detailed interviews and theory-driven coding of data. From this, a Rich Picture (Monk, Howard 1998) was developed to graphically present and relate stakeholder relationship information. The findings have implications for how stakeholder groups may be described, and how VGI can lead to enhanced products and services.

## 1. Introduction

Neogeography, the process of combining third party data with base maps to produce a mashup for a plethora of applications has - as Scharl and Tochtermann (2007) noted - had a profound impact on managing individual and organizational knowledge. While Neogeography may comprise entirely of professional information, a trend for including volunteered aspects arose around 2006, leading Goodchild (2007) to coin the term Volunteered Geographic Information to describe the creation of geographic information by largely untrained volunteers.

One of the key issues identified through analysing current research relative to the human issues of neogeography is *'how VGI maps and mashups are produced and used, and how do involved stakeholders interact'* (Rouse, Bergeron & Harris 2007, Harding et al. 2009). This paper illustrates how VGI is contributed and utilised by different individuals in terms of map choice, use of information, trust, influence, community, concerns, tensions, idealism and relationships.

## 2. Supporting Literature

The net of users connected to VGI, and those who have an influence on mashup design requirements may be considered stakeholders (Sommerville 2001). As Preece et al. (2002) demonstrated, the net of stakeholders is really quite wide, so it is useful to consider different categories of stakeholder, or as Coote and Rackham (2008) categorised: consumers, special interest [mapping] groups, local communities and professionals. Although VGI stakeholders may not be mutually exclusive (i.e. a stakeholder may be only a *consumer*, or also a *consumer* and also a *producer* of VGI) the simplified 'purist' model allows a more effective exploration of stakeholders than would otherwise be achievable in this paper.

The theoretical framework for this article is based on the unidimensional perspective of value to the stakeholder groups (Sweeney, Soutar 2001, Sheth, Newman & Gross 1991). This describes how consumer value is constructed from emotion, function, knowledge, legal and cost dimensions. This framework was selected due to its ability to demonstrate user perceptions that can influence adoption.

## 3. Methodology

This paper is based on detailed qualitative analysis of the *stakeholders* of VGI, focussing on OpenStreetMap, Ordnance Survey and 'Google My-Maps'. The study was conducted by interviewing lead users, discovered via '*snowball' non-probability* sampling (Robson 2002) through mailing lists, personal contacts and open advertisements. Topics covered included background ecology, involvement in VGI and intergroup interaction.

Interviews were recorded, transcribed and *'open coded'* (Robson 2002) under a thematic analysis approach (Aronson 1994, Braun, Clarke 2006) using NVivo 8 qualitative analysis software. Salient themes were collated to reflect topics delivered by the participants before being further reduced to best demonstrate interactions, concerns and perceptions of each stakeholder group (axial coding). This information was used to generate a Rich Picture by grouping respondents' transcriptions into stakeholder groups, and then drawing a 'Rich Picture' to graphically express the stand points of each stakeholder category. A Rich Picture provides a broad, high-grained view of the problem situation focusing on structure, process and concerns of the stakeholder (Monk, Howard 1998) – and is used to illustrate the main findings from this study.

**Table 1.** Breakdown of Study Respondents

| Mapping Project | *Consumers* | *SIMGs* | *LCs* | *VGI Professionals* | *PGI Professionals* |
|---|---|---|---|---|---|
| OpenStreetMap | | 6 | 1 | 1 | |
| Google My Maps | 3 | | | 1 | |
| Ordnance Survey | | | | | 2 |
| **TOTAL** | **3** | **6** | **1** | **2** | **2** |

While the number of respondents in this investigation was limited, the network of users who these respondents relate to is potentially quite wider. This causes the study to reflect on a much wider net of users than the total number of respondents.

## 4. Data Analysis & Results

Figure 1 identifies stakeholders alongside relationships, flow of information or influence (arrows), tensions (crossed swords) and concerns (thought bubbles) as described by Monk and Howard (1998).

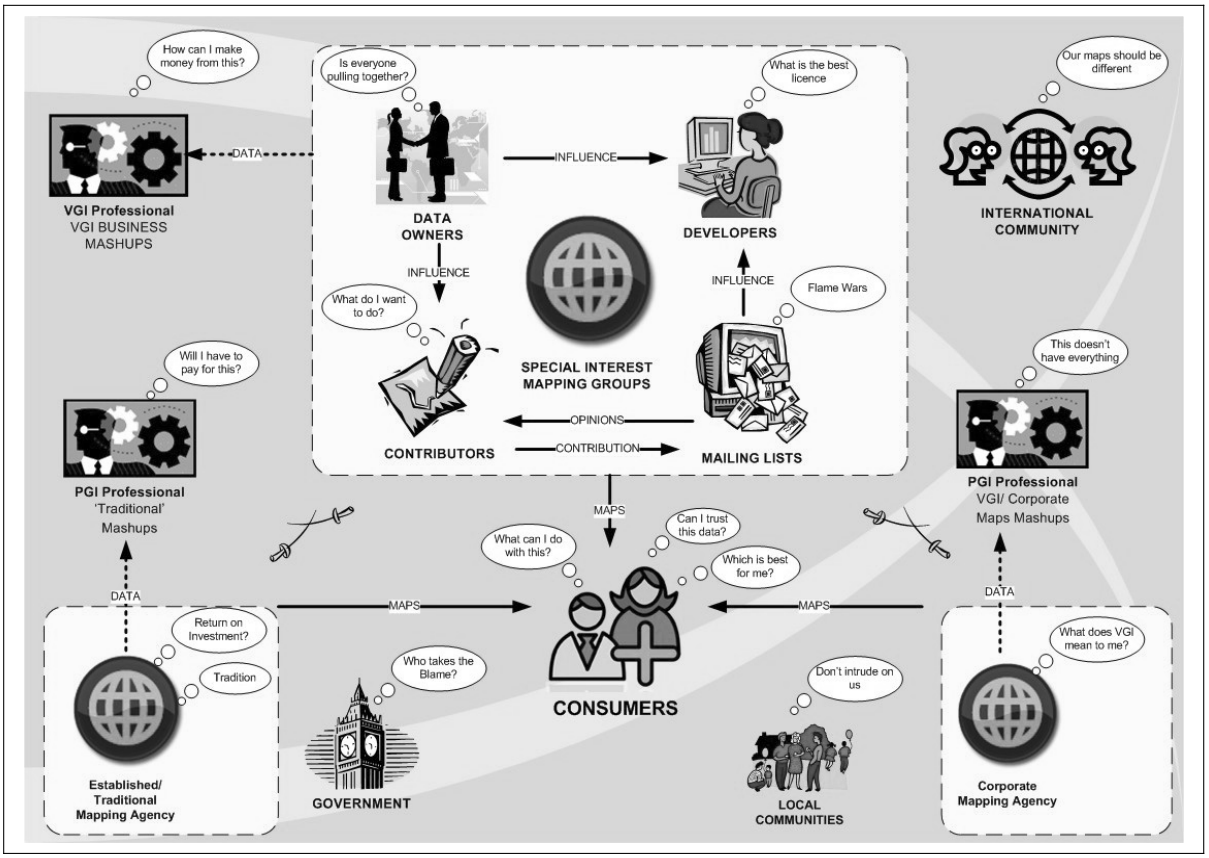



**Figure 1.** A Rich Picture of VGI Stakeholders

The following is a summary of the key points from the study.

*Consumers* select their map product to best fit their circumstances with little loyalty to the brand; as one respondent commented:

> *Apart from using it like everybody does in terms of looking for places and directions, I've used Google My Maps, at the moment mainly for my own use... I've used it in a work context because I was trying to organise a meeting.*

*Special Interest Mapping Groups* (SIMGs), *contributors* (SIMGCs) and *professionals* are particularly vested in the use of their groups' map; as one respondent commented:

> *I'll often check out to see if the local CTC has a website [same map project involved in] to see what's on there. And being able to find where the tea places are in the locality is quite useful.*

*Consumers* use their chosen map to fit their requirements within the usability boundaries set by the product's interface. *SIMGCs* produce data for group members and external parties to use their data; as one respondent commented:

> *It's mainly just a project to collect data... we hope other people will use it for whatever they feel free to use it for.*

*Professionals* aim to take the SIMGCs data and combine it with proprietary data as long as it enhances their business position.

*Consumers* are concerned about the accuracy of data within their map. *SIMGCs* are less concerned about inaccuracies as they have a vested interest in improving the data - seeing gaps as opportunities; as one respondent commented:

> *It has its faults but there are no glaring errors... It's very much if you don't like it you can fix it yourself which appeals to my, well, sense of working I suppose.*

*Professionals* are concerned about data validity and how inaccuracies may hurt their business position, as well as concern over what VGI actually means to their customers; as one respondent commented:

> *If I'm dispatching ambulances, and I know that I need to get to the patient within 7 minutes, can I trust the volunteer captured information?*

Although these reflections on user perception of accuracy may be suggested from this study, further research is required to produce more definite statements about the origin and in-depth user perception of VGI accuracy.

*Consumers* have little influence on their chosen product's development due to their position outside of the project, and return little or no data to the *SIMG*. *SIMGCs* have a direct and democratic influence on how data is produced and utilised within their project, shown by their position inside the *SIMG* and data transfer to other members via the mailing lists and wiki's. Professionals have limited or no influence on the production of the VGI they utilise; as one respondent commented:

> *All we can do is we can influence the direction this takes by offering suggestions*

*Consumers* are not necessarily part of any mapping related community. Both SIMGCs, professional and local communities have wide networks of associates with which they collaborate; as one respondent commented:

> *It's just a way of computer geek socialising at the end of the day*

*Consumers* displayed little or no tension between different map products during their interviews. *SIMGs* and *professionals* can be in constant tension with each other as their agendas and ideologies do not necessarily fit with each other.

> *It kind of annoys me that Google are potentially using the same kind of idea. [OpenStreetMap SIMGC]*

Between professional bodies, business rivalry may exist but they work alongside each other. Internally, *SIMGs* have many disputes over fundamental issues such as licensing, application of their VGI and future directions; as one respondent commented:

> *I will be chatting to my opposite number at Microsoft, and my opposite number at Google... we shouldn't even be friendly for Christ's sake according to the old fashioned rules of how you do business, and those old fashioned rules don't really apply any more.*

## 5. Discussion and Conclusions

The semi structured interviews and the subsequent production of the rich picture have highlighted the main stakeholders in VGI from a user-centred design perspective. The main outcome from this research has been that while stakeholders of VGI may often share common perceptions (e.g. *SIMGs*, *SIMGCs* and *professionals* having a vested in the use of their groups' map), often different

stakeholders will perceive elements of VGI very differently, based on which stakeholder group they may be identified with. The greater outcome of this study has been the examination of how and to what extent these similarities and differences occur.

Although this study was based upon value theory, determining a stakeholder collective perception of value is an elusive concept (Zeithaml 1988). However, if considering value as the improvement to a stakeholder's condition through utilising VGI (Menou 1995), then a salient increase in stakeholder value can be observed in all functional and work related perceptions.

The Rich Picture provided a visual framework to identify the interaction of stakeholders in terms of information flow between stakeholders; and inter-group tensions. The Rich Picture effectively provided context to the research outcomes and represented stakeholder relationships in an easily accessible fashion.

The implication of this work should be to provide a framework of VGI stakeholders to be utilised within future user-centred VGI research. This may take the form of deciding which stakeholder group(s) to target, or being able to relate their experiences to directly associated stakeholder groups. Additionally non user-centred GI research may be able to relate to the VGI end-user base in terms of which stakeholders may utilise their outcomes.

Future research will further investigate the multidimensional stakeholder perspectives on VGI for stakeholder groups, focusing on specific stakeholders and particular problem spaces.

## 6. References

Aronson, J. 1994, "A pragmatic view of thematic analysis", *The Qualitative Report,* vol. 2, no. 1, pp. 1-3.

Braun, V. & Clarke, V. 2006, "Using thematic analysis in psychology", *Psychology,* vol. 3, pp. 77-101.

Coote, A. & Rackham, L. 2008, "Neogeography Data Quality - is it an issue?", *AGI Geocommunity '08*Association for Geographic Information (AGI), http://www.agi.org.uk/SITE/UPLOAD/DOCUMENT/Events/AGI2008/Papers/AndyCoote.pdf.

Goodchild, M.F. 2007, "Citizens as Sensors: The world of Volunteered Geography", *GeoJournal,* vol. 69, no. 4, pp. 211-221.

Harding, J., Sharples, S., Haklay, M., Burnett, G., Dadashi, Y., Forrest, D., Maguire, M., Parker, C.J. & Ratcliff, L. 2009, "Usable geographic information – what does it mean to users?", *Proceedings of the AGI GeoCommunity '09 Conference*AGI GeoCommunity, http://www.agi.org.uk/SITE/UPLOAD/DOCUMENT/Events/AGI2009/papers/JennyHarding.pdf.

Menou, M.J. 1995, "The impact of information—II. Concepts of information and its value", *Information Processing and Management,* vol. 31, no. 4, pp. 479-490.

Monk, A. & Howard, S. 1998, "The Rich Picture: A Tool for Reasoning About Work Context", *Interactions,* vol. 5, no. 2, pp. 21-30.

Preece, J., Rogers, Y. & Sharpe, H. 2002, *Human-computer interaction,* John Wiley & Sons, United States of America.

Robson, C. 2002, *Real World Research,* Second edn, Blackwell Publishing, Oxford.

Rouse, L.J., Bergeron, S.J. & Harris, T.M. 2007, "Participating in the Geospatial Web: Collaborative Mapping, Social Networks and participatory GIS" in *The Geospatial Web: How Geobrowsers, Social*

*Software and the Web 2.0 are shaping the Network Societ*y, eds. A. Scharl & K. Tochtermann, Springer, , pp. 153-158.

Scharl, A. & Tochtermann, K. (eds) 2007, *The geospatial web: how geobrowsers, social software and the Web 2.0 are shaping the network societ*y, Springer Verlag.

Sheth, J.N., Newman, B.I. & Gross, B.L. 1991, "Why we buy what we buy: a theory of consumption values", *Journal of Business Research,* vol. 22, no. 2, pp. 159-170.

Sommerville, I. 2001, *Software engineering,* 6th edn, Addison-Wesley, Harlow, UK: Addison-Wesley.

Sweeney, J.C. & Soutar, G.N. 2001, "Consumer perceived value: The development of a multiple item scale", *Journal of Retailing,* vol. 77, no. 2, pp. 203-220.

Zeithaml, V.A. 1988, "Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence", *The Journal of Marketing,* vol. 52, no. 3, pp. 2-22.

## 7. Biography

*Christopher J. Parker is a second year PhD Research Student at Loughborough University, focusing on the stakeholder perceptions of volunteered geographic information from a human factors perspective. Andrew May and Val Mitchell are Research Fellows interested in user-centred design of new technologies.*

# Development of a server to manage a customised local version of OpenStreetMap in Ireland

Błażej Ciepłuch[1], Jianghua Zheng[1], Peter Mooney[1,2] , Adam C. Winstanley[1]

[1]Department of Computer Science, National University of Ireland Maynooth, Co. Kildare, Ireland
Tel. (+353 1 708 3847) Fax (+353 1 708 3848)
bciepluch@cs.nuim.ie (corresponding author)

[2] Environmental Research Center, Environmental Protection Agency, McComiskey House, Richview, Clonskeagh, Dublin 14. Ireland

KEYWORDS: OpenStreetMap, Geodata server, Tile Generation, Mapnik

## 1. Introduction

In this paper we describe the software architecture of a web-based GIS server system for the delivery of environmental research data in Ireland. The central component in this system is OpenStreetMap which provides the base layers of geographical data. An OpenStreetMap data collection campaign for our university town was carried out earlier this year yieldeding a spatially rich OpenStreetMap representation of Maynooth. Our server (OpenStreetMap database, supporting software, and specially generated map tiles) has been used by several GIS and location-based services projects in our department. One such example is a mobile device-based pedestrian navigation system is described in this paper. We describe some of the components of our server system. This includes a description of the management of the local copy of the OpenStreetMap database and the generation of sets of customised map tiles.

## 2. Choosing OpenStreetMap for this project.

One of the initial specifications of this research project was the development of a GIS data server. All software on the server must be free and open source software. A key function of the GIS server will be the provision of Web Map Services (WMS) and Location-based Services (LBS) for environmental research data in Ireland. The server must be capable of delivering different types of maps for different types of environmental data and information. This free and open source specification provided this project with the opportunity to with OpenStreetMap. At the beginning of this project we decided to become fully immersed in the 'world' of OpenStreetMap and volunteered geographic information (Haklay and Weber, 2008). We undertook an extensive OpenStreetMap mapping campaign of Maynooth. This involved mapping almost every geospatial feature in the locality: from university buildings, to housing estates, to public services, to traffic lights, to cemeteries. As data collection was underway other projects in our department began using the data. This showed us that the application of OpenStreetMap in our university would extend beyond our project. All data from this mapping campaign is now part of the OpenStreetMap dataset available to the whole world. Minor edits and additions are performed regularly, often on a weekly basis, from our research team. A new quality checked version of the entire OpenStreetMap database for Ireland is downloaded automatically every week and our local OpenStreetMap database is subsequently updated. In the next sections we describe the generation of sets of location specific map tiles for presenting environmental research data and information. As updates occur to our local OpenStreetMap database the sets of map tiles are regenerated as necessary.

## 3. Server Setup and Configuration

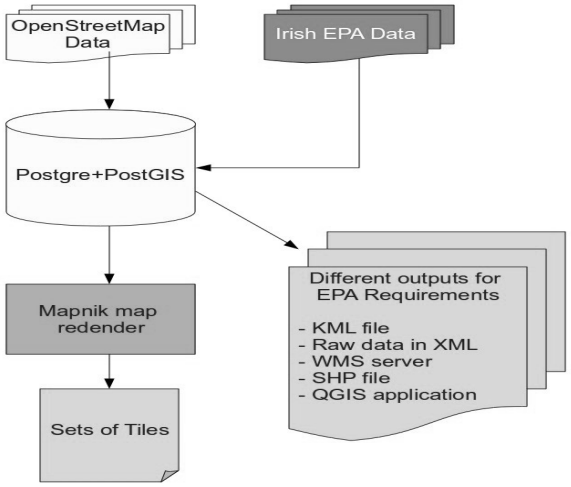



*Figure 1. Flowchart showing all software components of our server system*

Figure 1 shows a flow chart of the principal software components of our server system. Environmental data and information from the Environmental Protection Agency research is mostly in spreadsheets and database formats. Several PHP converter scripts have developed to import these data into the PostGIS database holding the OpenStreetMap data. Point based data was inserted into the OpenStreetMap point feature table while polygon data was inserted into the OpenStreetMap polygon tables. Mapnik software is used to generate sets of map tiles from the OpenStreetMap data. Using the GDAL ogr2ogr tool we can generate KML files and ESRI Shapefiles easily from the PostGIS database. Spatial output can be generated from entire tables within the database or as output from specific spatial queries specified in PostGIS SQL. From our local network we could provide access to members of our research team to access the local OpenStreetMap PostGIS database using QGIS. TileCache software is used to provide a Web Map Service (WMS) for access to the sets of map tiles.

## 4. Generation of customised map tiles

**I**n web-based GIS multiple layers of spatial information is usually presented by providing the user with a mapping base layer and a set of layers which can be "switched on and off" over the base map layer (Ciepłuch et al, 2009). This works well for many applications but can become unwieldy when there are a very large number of optional layers. This can also be awkward for users of web-based GIS applications on mobile devices. To present the results of environmental research in Ireland it was decided to generate sets of map tiles which embedded the environmental information and data into the map tiles to combine several layers into a single layer. For each set of map tiles certain map features were omitted, included, or had their cartographic representation customised depending on the environmental datasets. For example: for water related projects water features are emphasised on the maps while features such as roads, railways, etc. are removed or customised. The Mapnik (2009) application is used to generate sets of map tiles from PostGIS databases holding OpenStreetMap data. A key component of Mapnik is the XML file called the Mapnik stylesheet. In this configuration file are the rules for generation of tiles – what types of icon should represent point features, through what range of map scale should a certain feature be rendered, etc Rules are expressed in XML for any feature that is rendered on the tile and is represented in the database. Figure 2 shows an example of a rule from the Mapnik stylesheet where any polygons representing 'places of worship' are rendered in different colours depending on the zoom level the tile is requested at.

```
<Rule>
  <Filter>[amenity] = 'place_of_worship'</Filter>
  &maxscale_zoom10;
  &minscale_zoom14;
  <PolygonSymbolizer>
     <CssParameter name="fill">#777</CssParameter>
     <CssParameter name="fill-opacity">0.5</CssParameter>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <Filter>[amenity] = 'place_of_worship'</Filter>
  &maxscale_zoom15;
  <PolygonSymbolizer>
     <CssParameter name="fill">#777</CssParameter>
     <CssParameter name="fill-opacity">0.5</CssParameter>
  </PolygonSymbolizer>
  <LineSymbolizer>
    <CssParameter name="stroke">#111</CssParameter>
    <CssParameter name="stroke-width">0.3</CssParameter>
  </LineSymbolizer>
</Rule>
```

*Figure 2. A Mapnik style file rule controlling the rendering of places of worship at different zoom levels*

Figure 3 provides an example of where the customisation of the local version of the OpenStreetMap database and generation of customised map tiles. This set of map tiles show all map features in the Irish language. Irish is an official working language in Ireland and distribution of some information must be provided in both English and Irish. During the data collection campaign the Irish names of streets, buildings, etc were collected and assigned to their corresponding features in the OpenStreetMap database. The Mapnik stylefile is adapted to include a rule to ensure that only the "name:ga" attribute was used to render the text names of buildings, streets, areas, etc. To our knowledge this is one of the first examples of a fully Irish only web-based map.

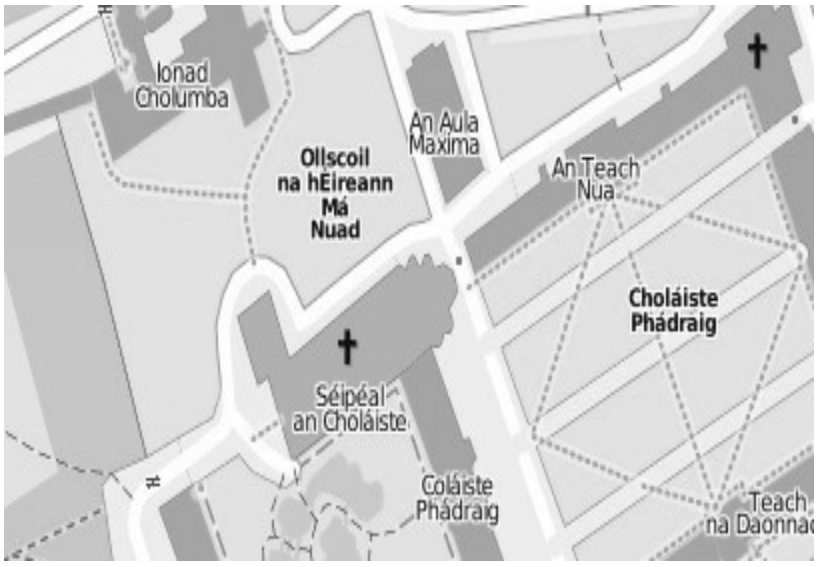



*Figure 3 OpenStreetMap for Maynooth showing Irish names for place names*

## 4. Accessing the customised map tiles

The sets of tiles generated by Mapnik are stored in different folders on the server hardisk. Tilecache is used to serve out these tiles to users who request them. To demonstrate the end-to-end functionality of the server: from input data, to customised tiles, to user interface we have built a web interface to the server. This is shown in Figure 4. Using the JavaScript library Open Layers the client can use the local tiles from our server but also access Web Map Services (WMS) from other remote locations. Two OpenStreetMap databases are available: our local version and the globally accessible OpenStreetMap database for Ireland. In Figure 4 the interface is shown where the user can click on the web-based map. All features matching their search criteria are returned and highlighted on the map. In the case of this example all places of worship are rendered red on the map if they lie within 200 meters of the map click location. The interface allows users to explore the spatial data in OpenStreetMap by providing a clickable map. Users can click on a location and explore the Point and Polygon features in the neighbourhood of the mouse click.

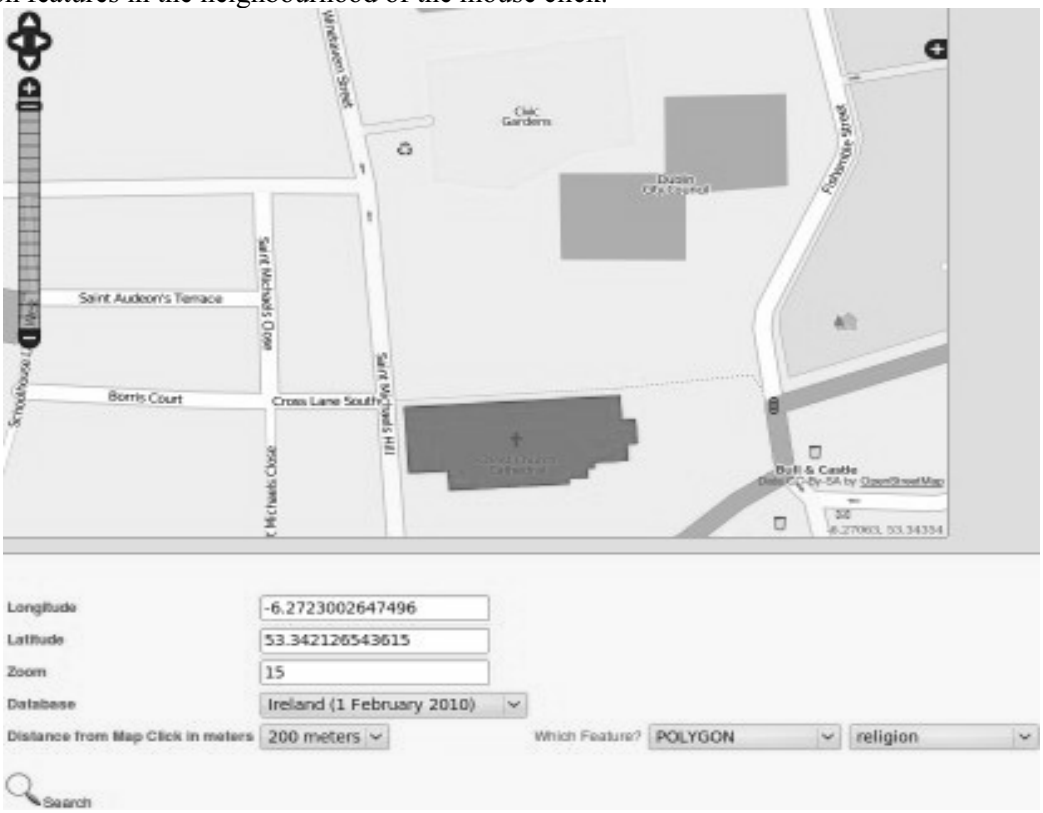



*Figure 4 A screenshot of the user interface to the server with customised tiles*

## 5. Applications accessing data and services from the local OpenStreetMap server

As mentioned above other projects in our department have begun to access the OpenStreetMap database and map tiles from our local server. Access to the PostGIS database and map tiles from applications residing on the same server or in the same network provides quicker query response time, better slippy map performance, and less overall network latency. One of the application using the server described in this paper is a Location-Based Service (LBS) for pedestrian navigation in the Maynooth university campus and Maynooth town. A special set of map tiles are generated which use darker colours for feature rendering which makes it easier for most mobile phone users to read when they are in bright environments. The local OpenStreetMap database is queried for spatial attribute information to supplement shortest paths generated specifically for pedestrians. Spatial attribute information include: Irish language placenames, historical information, and postal address details. Figure 5 shows a screenshot of a mobile device accessing the local server with OpenStreetMap database access and customised local tiles.

*Figure 5. Mobile client using both the local tiles and OpenStreetMap database*

## 5. Conclusions

The paper describes work that is very much in progress. There is an incorrect perception that the setup and configuration of a customised OpenStreetMap data and tile server is a very difficult task. Any perceived difficulties in setup and configuration are offset by the advantages of having a very spatially rich database of geographical information for our locality available on a GIS server system. The next milestone is to prepare sets of tiles based on the likely profiles of potential users of environmental research data and information. This will involve generating tiles holding a common environmental theme but are different for urban and rural areas due to different perceptions of the environment from citizens living in these areas. We are commencing research into inclusion of vernacular geographies into our local version of OpenStreetMap to include the local names of streets, buildings, and other geographical features different to those appearing in official gazetteers.

## Acknowledgements

The authors gratefully acknowledge the funding for this research provided by the Environmental Protection Agency Ireland's STRIVE programme. The work is part of a larger project Geoinformatics Services for Improved Access to Environmental Data and Information (2008-FS-DM-14-S4) is a 5 year project from 2008 – 2013. Dr. Mooney is PI for this project. Dr. Adam Winstanley is CO-PI for the Location-based Services strand of the Science Foundation Ireland funded STRAT-AG programme. The support of STRAT-AG is gratefully acknowledged also.

## References

Ciepluch, B., Mooney, P., Jacob, R. and Winstanley, A. C., 2009. Using openstreetmap to deliver location-based environmental information in ireland. SIGSPATIAL Special 1(3), pp. 17–22.

Haklay, M. M. and Weber, P., 2008. Openstreetmap: User generated street maps. IEEE Pervasive Computing 7(4), pp. 12–18.

Zheng, J., Ciepluch, B., Mooney, P., Jacob, R. and Winstanley, A. C Location Based Services of University Town Based on OpenStreetMap: NUI Maynooth as an example CIICT 2009 . Proceedings of the China-Ireland Information and Communications Technologies Conference

## Biography

*Błażej Ciepłuch is a 2nd year PhD student in Computer Science. He obtained an MSc in Computer Science from Poznań Technical University in Poland in 2006. Dr. Jianghua Zheng is researcher in pedestrian navigation project. Dr. Peter Mooney is Błażej's PhD supervisor and is an environmental research data manager with the Irish EPA. Dr. Adam Winstanley is head of the Computer Science Department at NUIM and is a CO-PI of the LBS Strand of the STRAT-AG project.*

# Polygon Processing on OpenStreetMap XML Data

Fangli Ying, Peter Mooney, Padraig Corcoran and Adam C. Winstanley

[1]Department of Computer Science, National University of Ireland Maynooth, Co. Kildare. Ireland
Tel. +353 1 62801011
Email: fying@cs.nuim.ie, peter.mooney@nuim.ie, padraigc@cs.nuim.ie adamw@cs.nuim.ie

KEYWORDS: OpenStreetMap, polygon extraction and analysis

## *1. Introduction*

One of the many ways of accessing the raw geographic data collected and distributed by OpenStreetMap is by using the OpenStreetMap (OSM) XML data format. The OSM XML can be explored using standard XML visualisation and search tools. This paper describes the development of a software tool for the examination of polygons represented in OpenStreetMap XML. Given an OSM XML file corresponding to a specific geographical area the software performs the following two functions during polygon examination:

1. *Examination of the connectedness of water features*. The OSM XML is checked to ensure that water features are correctly connected and are consistent with the physical reality. For example a river or stream flowing into a lake. Potential spatial connectivity problems are highlighted and can be used for quality control purposes for the OpenStreetMap data.
2. *Automated identification and selection of suitable sets of polygons as input for testing generalisation algorithms*. This software automatically identifies suitable sets of polygons for testing generalisation algorithms by calculating the overall complexity for each polygon in a specified geographical region. If this overall complexity is high then all of the features are extracted from the OSM XML and output in ASCII format to the generalisation algorithms.

The software is written in Java. Figure 1 provides a schematic of the individual components in the software tool. An OSM XML file is presented as input. Firstly connection issues in the spatial data are identified and a report is generated on these issues. An OSM community member can then use this information to update the OSM database to address these issues. Secondly the tool computes a number of polygon characteristics for each of the polygons in the OSM XML file. If the set of polygons are identified as sufficiently complex they are reformatted into ASCII file format where they are passed to the generalisation module. This is a separate software tool that runs a generalisation algorithm on this spatial region and is not described in this paper. A local database copy of OSM is then updated automatically with the generalised representations of the polygons inserted. The polygons in the OSM XML are captured in a number of ways: from GPS data collections, tracing over aerial imagery, and bulk upload of spatial data to OpenStreetMap.

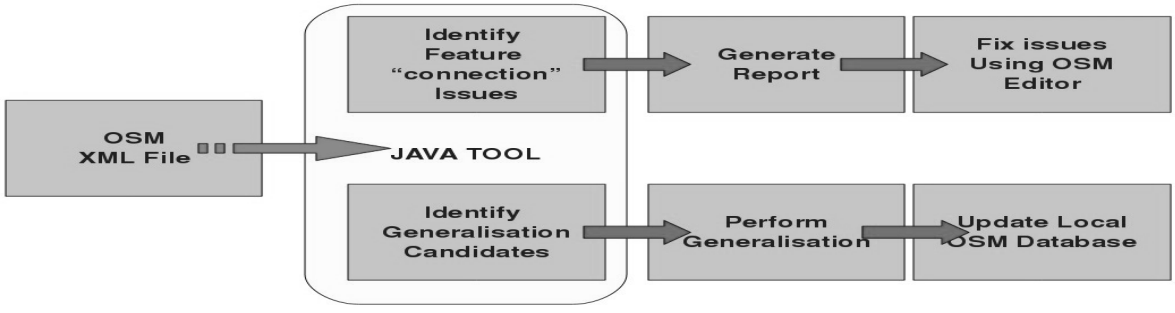



**Figure 1.** Schematic Diagram showing components of the software tool for polygon processing

## 2. Spatial Representation in OpenStreetMap XML

While the OSM XML schema is easy to understand, it is very difficult to identify geographic features which are connected to each other without visualising the XML as a tree-based structure or rendering the data as a set of base map tiles or an overlay on another map. The problem is particularly difficult when the OSM XML represents (1) a very large geographical area, and/or (2) contains a very large number of lines and polygons. OSM XML contains points, lines, and polygons. Every spatial attribute (or *tag*) corresponding to each point, line, or polygon feature is included in the XML. It was a specific requirement of this software to operate directly on the OSM XML by downloading OSM XML *on-the-fly* from the Internet and performing the processing in real-time. Consequently, there is no need for setup of spatial databases or desktop GIS. Data in OpenStreetMap is contributed by members of the OSM community. In the OSM database there are many features, sometimes simple polygons, which are often greatly over-represented when considering them for applications such as Location-based Services (LBS). This means that the GPS traces uploaded to OSM for these features were sampled at a very high rate. Regular shaped polygons (squares, rectangles, triangles) representing buildings, car parks etc often contain many hundreds of points. More complex polygon shapes, such as those representing natural features such as lakes are often represented by many thousands of points. Given the characteristics of the polygon this is often necessary. However in some cases the number of points used to represent the polygon could be reduced. Relationships between points and polygon/line features are represented simply in OSM XML. Each point in OSM has a unique ID (OSM_ID). Each polygon/line in OSM also has a unique ID (OSM_ID). In the OSM XML each polygon/line is represented as a <way> with a unique OSM_ID. Each <way> is a ordered collection of <node> features with unique OSM_ID.

## 3. Examination of the connectedness of water features in OpenStreetMap.

Given a feature class the software automatically identifies polygons (belonging to that particular feature class ie. waterbodies, forest and woodland, etc) in the area specified in the OSM XML file. The software identifies all other line and polygon feature intersections. Based on specified rules for spatial feature connectivity (river flows into lake, steam flows into river) it determines features not correctly connected within the OSM XML. Figure 2 shows an example from the OpenStreetMap database correct as of Feb 25th 2010. Lough Erne in County Fermanagh, Northern Ireland is shown. Lough Erne (the Upper and Lower loughs) are actually widened sections of the River Erne. The representation in OpenStreetMap is incorrect where the two loughs are represented as two distinct unconnected features. The physical reality is that Upper lough flows into the Lower lough. For water features our software computes the nearest nodes between pairs of polygons. If it is found that this distance is less than a pre-determined value we then infer that these features should be physically connected. For example a river or stream separated by less than 20 meters from a lake. Potential errors in representation are then reported to allow OpenStreetMap contributors to address the issues. For natural landscape features such as rivers and lakes this software could assist in improving the quality of the OpenStreetMap database by suggesting improvements in terms of feature connectivity.

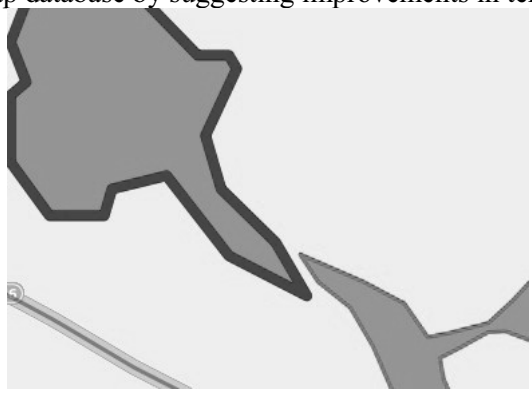


**Figure 2.** Visual Representation of apparently connected polygon features

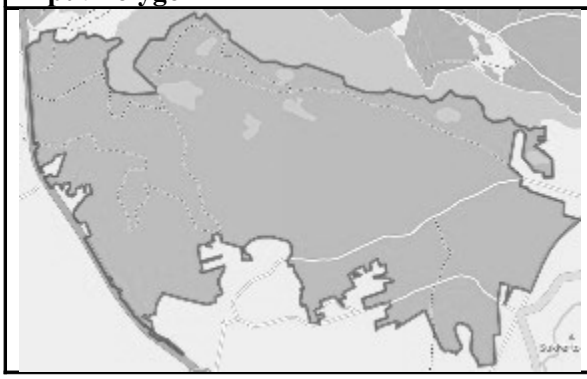
## 4. Automated identification and selection of suitable sets polygons as input for testing generalisation algorithms.

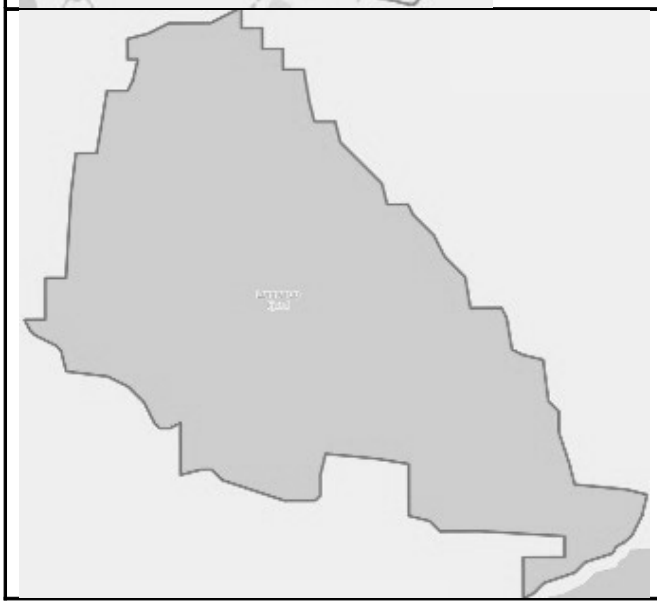
In our software we have implemented a strategy to identify suitable sets of polygons for use as input to testing generalisation algorithms. The software attempts to calculate characteristics of each polygon which can be used to give an overall measurement of the complexity of polygon. Four principal characteristics are outlined as follows which together can help to automatically describe the complexity of a polygon without human visual evaluation of the complexity (Bryson and Mobolurin 2000). The software attempts to estimate the polygon complexity in the smallest number of steps which are outlined as follows:

1. **N:** The number of points representing the polygon
2. **Turning angle (k)** at each node. The overall measure of turning angle is the mean of all k values denoted as K. The greater the variation in turning angle (large K values) the more complex the polygon is likely to be (Latecki and Lakmper,1999)
3. **Cicularity:** Simple circularity measure - ratio of the square of the perimeter length to the area (normalised between 0.0 and 1.0)
4. **Area Ratio:** Convex Hull circularity – normalised ratio of difference between area of polygon and its convex hull. Larger ratios (close to 1.0) can indicate a complex polygon.

Other measurements are calculated including: distribution of normalised distances between adjacent nodes in the polygon, width of the polygon, and distribution of normalised distances between every vertex the centroid of the polygon. This purpose of the steps above is to allows us to select suitable test data as input. Most of the time only a subset of the steps will be required for the software to check if the polygon is complex or not (Brinkhoff *et al*, 1995). Table 1 below shows some examples of the output from our software for some singular polygons input.

**Table 1.** Example of complexity analysis results for three sample polygons.

| Input Polygon | Class | Rationale |
|---|---|---|
|  | **Complex** | **Generalise = Yes:** 559 nodes, Circularity = 0.285, Area Ratio = 0.255. K = 0.0018. Very low K value means that this shape could be generalised as some vertices could be removed without dramatically altering the structure of the polygon. |

| | | |
|---|---|---|
|  | **Simple** | **Generalise = No:** 40 nodes, Circularity = 0.246, Area Ratio = 0.097. K = 0.0256. This polygon is very circular. There is a small number of nodes. The K value is very high meaning that all nodes have very high significance. |
|  | **Simple** | **Generalise = Yes:** 177 Nodes, Circularity = 0.311, Area Ratio = 0.251. K = 0.0057. The K value for the polygon is very low meaning that many insignificant nodes exist in the polygon. These could be removed by generalisation. |

## *5 Conclusions*

The software tool described has been developed to (1) identify spatial data quality issues for connected polygon features in OpenStreetMap and (2) select suitable sets of polygons from OpenStreetMap for use in the testing of generalisation algorithms. The software has been developed to work on small geographical regions – approximately 10Km2. It is guided by the user who provides specifies the region that may have problems in relation to the connectivity of polygon features. This software will assist the authors in identifying these connectivity issues in the OpenStreetMap database. Problems encountered can be fixed using one of the many OpenStreetMap data editors available. Automated identification of polygons for testing generalisation algorithms is a very interesting topic and the algorithmic approach described is at an early stage. Human visual perception can identify complex polygons very quickly when presented with a cartographic representation of the polygons (Brinkhoff *et al*, 1995). Our software tool aims to identify suitable generalisation candidates quickly from the OSM XML provided as input by calculating a suitable subset of the polygon characteristics outlined above. Several characteristics of the polygons must be calculated in order to automatically assess the complexity. For applications such as Location-based Services (LBS) the spatial data in the OpenStreetMap database is one where generalisation will always be needed as new polygon features are added and existing polygon features are edited. We will continue to investigate other methods for shape complexity identification particularly from the fields of computer vision and shape recognition. This paper does not provide details on validation of the results from the generalisation algorithms employed. This work is described in other papers by the authors.

## Acknowledgements

Fangli Ying is funded by Irish Universities Association and China Scholarship Council. Peter Mooney is funded by the STRIVE Research Programme administered by the Environmental Protection Agency of Ireland. All of the authors are part of the Location-based Services (LBS) Strand under Science Foundation Ireland's STRAT-AG funding at the National Center for Geocomputation (NCG) at NUI Maynooth. Adam Winstanley is co-PI of the LBS Strand.

## References

T. Brinkhoff, H.-P. Kriegel, R. Shneider, A. Braun, *Measuring the complexity of spatial objects*, Proceedings of the 3rd ACM International Workshop on Advances in Geographic Information Systems, Baltimore, MD, 1995, pp. 109–117

N. (Kweku-Muata) Bryson, A. Mobolurin, *Towards modeling the query processing relevant shape complexity of 2D polygonal spatial objects*, Information and Software Technology, Volume 42, Issue 5, 1 April 2000, Pages 357-365

L. J. Latecki and R. Lakmper, *Convexity rule for shape decomposition based on discrete contour evolution*, Computer Vision and Image Understanding, vol. 73, no. 3, pp. 441 – 454, 1999.

## Biography

*Fangli Ying is a first year PhD student at the Department of Computer Science. He holds a degree in Computer Science. He commenced his PhD in October 2009. Peter Mooney is a postdoctoral research fellow at the Department of Computer Science and at the Environmental Protection Agency of Ireland. Padraig Corcoran is a lecturer at the Department of Computer Science with research interests in machine learning, pattern matching, and computer vision. Adam Winstanley is head of the Computer Science Department and is co-PI for the Location-based Services strand of the Science Foundation Ireland STRAT-AG project at the National Center for Geocomputation at NUIM*

# Rate-my-place: a social network application for crowd-sourcing vernacular geographic areas

Julian Rosser[1, now at: 2], Jeremy Morley[2]

[1] Department of Civil, Environmental & Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, United Kingdom.

[2] Centre for Geospatial Science, University of Nottingham, NG7 2RD, United Kingdom

KEYWORDS: vernacular geography, web 2.0, crowd-sourcing

## 1. Introduction

Web and mobile applications utilising spatial data are becoming more commonplace, and increasingly consumers of geospatial products are found outside professional geographically related fields. In spite of this growing geographic awareness, most individuals use vague and informal terms when referring to places. Such places are identified here as vernacular areas and can remain unrecognised in the official designations of spaces usually used to process geographic information such as gazetteers. Furthermore, the places that are recognised may be perceived differently by individuals in comparison to the official definition.

The growth of the Internet and the move toward context-aware computer systems present new challenges as to how spatial information is accessed. It has been estimated that between 13-15% of search engine queries contain a place name (Sanderson & Kohler, 2003), and it is recognised that Geographic Information Retrieval (GIR) will improve with added understanding of vernacular areas and terminology (Hill, 2006; Jones & Purves, 2008). Furthermore, integration of this information into GIS means more natural interfaces can be developed that incorporate vague spatial and non-spatial concepts (e.g. Cai et al. 2005). Traditionally, the methodology for deriving perceptions of areas has involved the use of sketch maps (see Montello, et al. (2003); Matei, et al. (2001)). More recently work has focused on mining the web for location data which may then be used to form boundaries both at small (regional) and large (neighbourhood) scales (see Arampatzis, et al., 2006; Pasley, et al., 2006).

Web 2.0 and technological advances in computer systems facilitate direct capture of the vernacular perceptions of many individuals and this paper reports on a pilot study for software capable of achieving this. The software is a social networking application, dubbed “Rate-my-place”, in which users are encouraged to rate and mark the extent of places on a map. After describing the design and implementation of the software, preliminary results of deploying the application online are presented. This is an ongoing project and illustrates work completed to-date. Over the coming months the software will be online more or less continuously promising further results from a much larger dataset of vernacular perceptions.

## 2. Conceptual design

User experiences of curiosity, fun and enjoyment have been found to form important motivations for the use and adding of applications to Facebook profiles (Hart, et al., 2008). So, in an approach similar to that used by Evans and Waters (2008), a spray can tool provides a novel and fun means of marking vague areas. Social networking environments such as Facebook have the advantage that small applications similar to Rate-My-Place are common and spread easily by word-of-mouth through the system, increasing the number of users accessing the software and hence the number of extents

contributed. While a single user may only remain interested in the application long enough to mark a few areas, widespread dissemination of the system ensures steady growth of the vernacular dataset.

The users of the application get to see the combined rating surface of their area of interest based on all contributions made to the system. The second, research aspect, and primary reason for undertaking this work, is to store the individual extents for aggregation based on common submitted place names. The vernacular areas derived for any name are fuzzy, matching the perceptual fuzziness of the boundaries. The collaboration is appropriate as it is the general public that are closest to being domain experts in the definition of informal areas. Rosch (1978) found that some objects make better examples of categories than others. Here, users identify the locations that are more typical to a particular region.

The rated areas must be combined together to form a final surface of the world that is meaningful to a regular user of the application. So, averaging the ratings as they are added into the system seems the most appropriate method for combination. For vernacular area generation, individual extents for a place name are summed together to map the level of agreement between contributors. This forms a separate processing task carried out independently of the rating surface composition.

Two distinct methods to simulate the spray can effect can be implemented in software, identified here as dithering and Gaussian patterns (see Figure 1). Testing of both mechanisms identified that users preferred the Gaussian spray, finding it easier to control than the dithering pattern.

**Figure 1.** Dithering and Gaussian spray patterns.

### 3. Implementation

Developing a spray effect for use in web browsers and processing the areas delineated by the tool poses several challenges. Firstly, on the client side a user's machine must be capable of rapidly rendering both map and spray effect. Users of social networking sites will not persist with an application that runs slowly or requires significant loading time. Secondly, the server must also rapidly process each spray extent so the user may view their contribution soon after its submission.

### 3.1 Client

Adobe Flash was chosen for its common availability, rendering performance and consistent display. The interface uses Google Maps and enables users to mark out areas with the spray can tool based on verbal ratings of "love", "like", "dislike" and "hate" (see Figure 2). The rating dictates the spray colour. A free text box collects an associated place name for each marked extent. It was thought that leaving the user to freely choose their place name might encourage contribution of vernacular terms which may be previously unrecorded. Furthermore, using a text box leaves the user free to qualify areas by spatial relationships if they wish, which is recognised as a way to identify locations (Jones, et al., 2003). The amount of base map detail provided to the user affects their area perception and so marking of extents was restricted to zoom levels 13 and 14. This provided appropriate detail for collection of vernacular perceptions of local areas and neighbourhoods in addition to a consistent scale for resultant extents. The map tool uses Facebook via the social networking site's API. Upon submission of three areas, names and ratings, users may opt to view the overall rating map.

The rating is kept deliberately open to allow users to interpret the ratings themselves. As research data we are interested in the individual sprayed area with its associated name from the text box and are less concerned in this project with the actual ratings. By combining sprayed areas with synonymous names we can explore the fuzzy area associated with each place name. The user interface on the other hand presents the mean rating surface so the user can see his/her rating in the context of an average view.

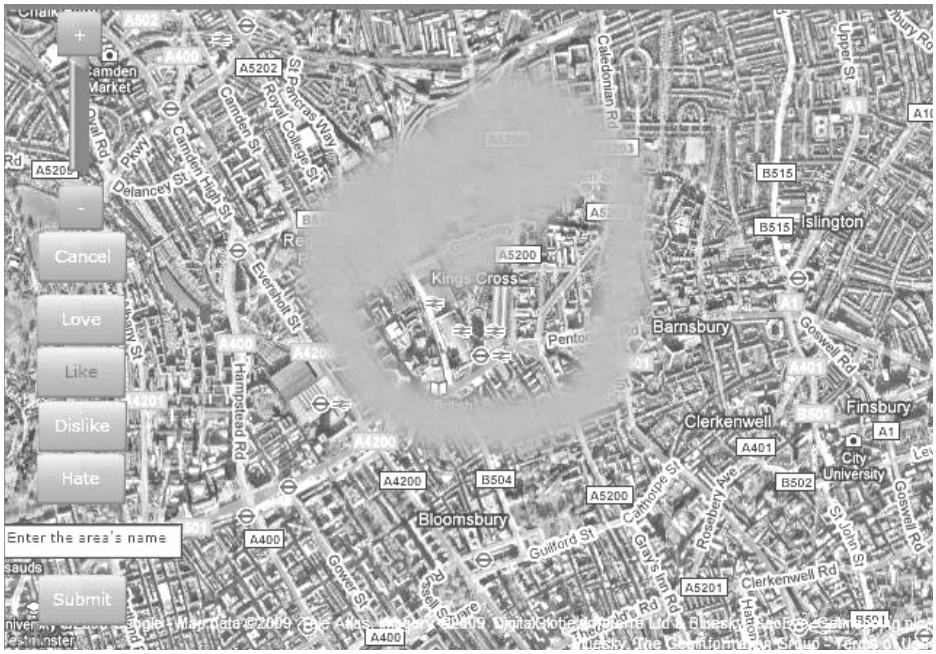



**Figure 2.** Marking an area using the spray tool.
©2009 Google – Imagery ©2009 DigitalGlobe, Infoterra Ltd & Bluesky, GeoEye, GetMapping plc, TheGeoinformation Group, Map data ©2009 TeleAtlas.

### 3.2 Server

The server-based component of the application comprises two parts: an average rating surface and vernacular database. After a user submits an extent it is added in real time to the average rating surface (cell size of 100m). The extent submitted by the user is first scaled according to their rating (see Table 1). The average (mean) value for each location is calculated through use of an additional raster which stores a count of the number of times each cell has been rated. A tiling program uses the rating raster to produce PNG tiles suitable for display as a web mapping overlay.

| Rating | Scaling range (min -> max) |
|---|---|
| Love | 0.5 -> 1.0 |
| Like | 0 -> 0.5 |
| Dislike | 0 -> -0.5 |
| Hate | -0.5 ->-1 |

**Table 1.** Scaling of cell values based on rating

Individual extents and their associated place names are also stored within a database which may be queried in order to extract and generate aggregated vernacular areas. The database stores the map extent and zoom level used during the capture process together with the string of points which composes the spray line.

## 4. Results

As a test run of the application, preliminary results were generated based on an online deployment over a five day period. Participants were asked to focus on rating areas of London. 127 place extents were submitted by 24 participants. Of these, 80% of named areas were distinct, meaning that 20% of areas had at least two individual perceptions (areal extents tagged with the same name). No disambiguation was attempted on the data, and so the number of duplicate extents is actually higher after consideration of spelling mistakes and abbreviations (e.g. “Oxford Street” and “Oxford St” are the same location but were identified as distinct places).

To illustrate the derivation of a vernacular area based on a number of spatial perceptions, five raster extents tagged as "Soho" were extracted from the database, summed and normalised to within a range of 0 and 1 (see Figure 3). Areas of strong agreement about the core location of the neighbourhood are clear, centred over Soho Square. As Soho does not have an official boundary for comparison, a Flickr Shapefile (based on the alpha shape of photo tags of Soho, London) is also shown (Flickr Shapefile, 2008). Obviously, Flickr Shapefiles have their own set of issues and restrictions; however, it is useful and interesting to see the boundary for comparison.

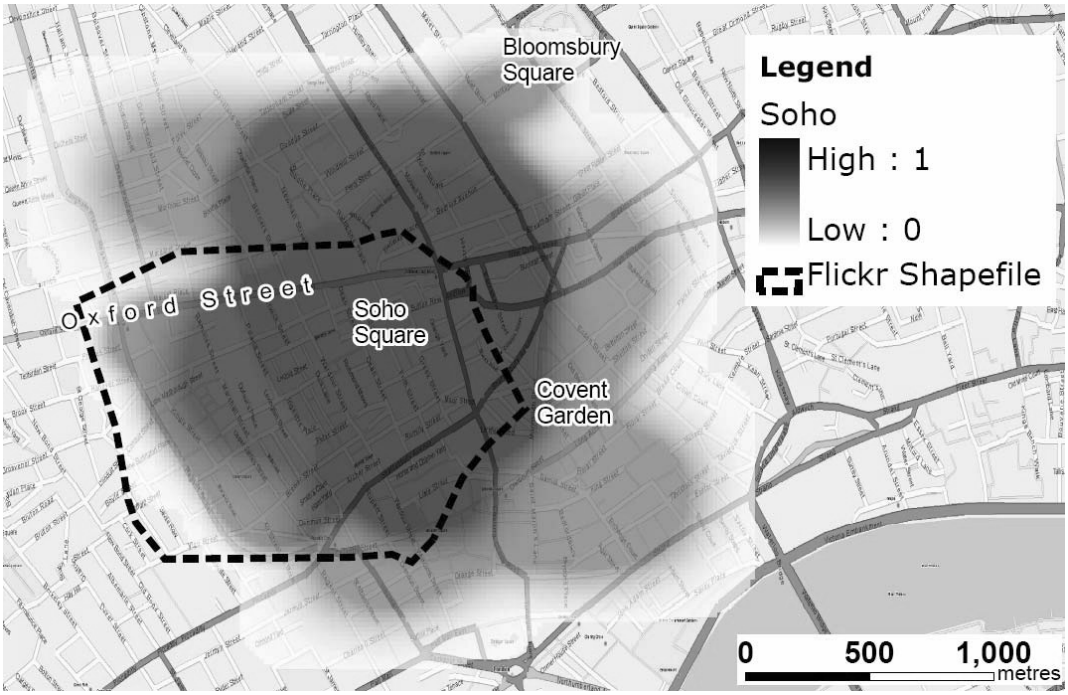



**Figure 3.** Area of Soho based on five submitted extents.
The map labels have been added here to provide additional context to the reader.


## 5. Conclusion

Work is underway to perfect aspects of the software and deploy the system for a longer period; but the results demonstrate the application as a useful mechanism for capturing vernacular geographic entities. Ongoing collection of data would provide a significant source of vernacular information, perhaps for comparison with a gazetteer, and add to that derived using other methods such as web mining. For end users, the integration of these geographies into products, for example route finding devices, would permit a more natural interaction for the operator. Similarly, local search queries would return more relevant records.

There are limitations to this approach of crowd-sourcing vernacular information. The user's focus on rating areas with a spray can tool is a fundamental concept to the work and the impetus for their use of the software. The elicitation of vague boundaries is the central aim of the application, yet it is ratings of areas that the user considers. Users may group neighbourhoods together in one area in order to assign the same rating. It is also unclear to what extent the individual's context may affect their boundary perception. The user's focus is on rating areas and so representations may be distorted e.g. "like" extents may be larger than "dislike". Furthermore, a user's reason for prescribing a particular rating to an extent is not recorded which may further influence their concept of place. As the system is revised, we hope to record aspects of physical context including location (provided via user's Facebook profiles and/or geo-referenced IP addresses).

Work may be pursued investigating the generation of a new extent where vernacular perceptions may be "crisped". Calculating the statistical significance of agreement offers one potential avenue for achieving this. Applied to a similar issue of vagueness found in use of spatial prepositions, Hall and Jones (2009) investigated active contours for crisping a field-based representation which might offer an alternative approach.

## 6. Acknowledgements

This work is supported by Ordnance Survey Research and the Centre for Geospatial Science at the University of Nottingham. The authors would like to thank the Ordnance Survey Research department for their valuable feedback and three anonymous reviewers for their helpful comments.

## References

Arampatzis, A., Van Kreveld, M., Reinbacher, I., Jones, C., Vaid, S., Clough, P, Joho, H. & Sanderson, M. 2006, Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems.* 30, pp. 436–459.

Cai G, Wang H, MacEachren A M and Fuhrmann S. 2005. Natural Conversational Interfaces to Geospatial Databases. *Transactions in GIS*. 9(2): 199-221

Evans, A.J. and Waters, T. 2008. Mapping vernacular geography: web-based GIS tools for capturing "fuzzy" or "vague" entities. *International Journal of Technology, Policy and Management.* Volume 7 (2), 1468 - 4322.

Flickr Shapefile. 2008. Flickr Shapefiles Dataset – Areas tagged Soho London. Available from: http://farm4.static.flickr.com/3024/shapefiles/43201_20081112_c63e66761e.tar.gz [Accessed 26 August 2009]

Hall, M. & Jones, C. 2009. Initialising and Terminating Active Contours for Vague Field Crisping. *GISRUK 2009*, pp. 395-397.

Hart, J., Ridley, C., Taher, F., Sas, C. & Dix, A. 2008. Exploring the Facebook Experience: A New Approach to Usability. *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, October 20-22, 2008, Lund, Sweden. Pp. 471-474.

Hill, L.L., 2006. *Georeferencing: The Geographic Associations of Information*. Cambridge, MA: MIT Press.

Jones, C.B., A.I. Abdelmoty, and G. Fu., 2003. Maintaining ontologies for geographical information retrieval on the web. In *On The Move to Meaningful Internet Systems 2003*, ODBASE'03.LNCS 2888, 934-51.

Jones, C.B., Purves, R.S. 2008. Geographical information retrieval. *International Journal of Geographical Information Science.* 22(3) pp.219-228.

Matei, S., Ball-Rokeach, S., & Qiu, J. (2001). Fear and Misperception of Los Angeles Urban Space: a Spatial-statistical Study of Communication-shaped Mental Maps. *Communication Research*, 28(4), pp. 429-463.

Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P. 2003. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*. 3(2):185-204. P.201.

Pasley, R., Clough, P. and Sanderson, M. 2007. Geo-Tagging for Imprecise Regions of Different Sizes. *Proceedings of Workshop on Geographic Information Retrieval GIR'07.*
Rosch, E. 1978. Principles of categorization. In: E. Rosch & B. Lloyd, eds., *Cognition and Categorization*. Hillsdale, N.J.: Erlbaum Associates. 27-48.

Sanderson, M. & Kohler, J., 2004, Analyzing geographic queries. In *Proceedings of the 2004 Workshop on Geographic Information Retrieval,* 29 July 2004, Sheffield, UK.

**Biography**

*Julian Rosser previously studied Geographic Information Science (M.Sc.) at University College London. He now works at the Centre for Geospatial Science, University of Nottingham. His primary interests are Web 2.0 and environmental GIS.*

*Jeremy Morley is Deputy Director of the Centre for Geospatial Science at the University of Nottingham, and an Honorary Senior Lecturer at UCL. His interests lie in Internet-based GIS and interoperability; and environmental modelling in GIS.*

# Uncertainty in the 2001 Output Area Classification for the census of England and Wales

Peter Fisher[1]

[1]Department of Geography, University of Leicester
Leicester, LE2 3TD, United Kingdom
Email: pff1@le.ac.uk

KEYWORDS: Geodemographics, Output Area Classification, 2001 Census, fuzzy classification, possibilistic classification

## 1. Introduction

The Output Area Classification (OAC) developed by Vickers and Rees (2006, 2007) presents a geodemographic analysis derived from data collected by the 2001 Census of England and Wales. As such it is a direct competitor for commercial products such as Acorn and Mosaic (Harris et al., 2005). Unlike its commercial competition, however, the OAC has been published together with the uncertainty of every Output Area (OA) being a member of the class to which it has been assigned.

The original classification used the well known *k*-means classification method (Vickers and Rees, 2007). It identified 7 Supergroups at the highest level of classification, 21 Groups at the next level and 52 Subgroups at the lowest level. Following analysis of the values of the variables used in the classification the seven classes at the highest level are labelled as: Blue Collar Communities, City Living, Countryside, Prospering Suburbs, Constrained-by-Circumstances, Typical Traits, and Multicultural Communities. Such descriptions indicate the vagueness of the classes derived. The names are interpretations of what could be the significance of the clusters based on within-class mean values for the variables used.

## 2. Fuzzy Classification

Fuzzy classification has been with us for a number of years, particularly since the ground-breaking work on Bezdek (1981; Bezdek et al., 1984). Instead of returning a single class for each case or observation as do classification algorithms based on Boolean sets, a fuzzy classification returns a degree to which every case belongs to every class. So instead of a one dimensional matrix resulting from the classification, a matrix results which has dimensions $c \times N$ where $c$ is the number of cases, and $N$ is the number of objects. Bezdek et al. (1984) introduced the Fuzzy *c*-means based on the well known *k*-means clustering method used for the OAC. It has been used and extended by a number of researchers including to the fuzzy *k*-means (Burrough and McDonell, 1998).

### 2.1. Fuzzy k-means

Burrough et al. (2000; Burrough ad McDonnell, 1998) present the fuzzy *k*-means as a variant of fuzzy *c*-means. In this method the fuzzy membership of ay class is related to the distance to the class centroid in the feature space by equation 1.

$$\mu_{fuzz(ij)} = \frac{\left[\left(d_{ij}\right)^2\right]^{\frac{-1}{m-1}}}{\sum_{l=1}^{k}\left[\left(d_{il}\right)^2\right]^{\frac{-1}{m-1}}} \quad \text{Equation 1}$$

Where $\mu_{fuzz(ic)}$ is the fuzzy membership of an object $i$ in the set or class $c$,
$d_{ic}$ is the distance of the object $i$ from the prototype of class $c$,
$m$ is the index of fuzziness, and
$c$ is the number of classes being considered.

This is subject to the conditions that:

$\mu_{ij} \in [0,1]$ for all $i$ and $j$ Equation 2

$0 < \sum_{j=1}^{N} \mu_{ij} < N$ for all $i$, and Equation 3

$\sum_{l=1}^{k} \mu_{il} = 1$ for all $i$. Equation 4

where $N$ is the number of objects.

The last condition (Equation 4) may be unrealistic, however. There is no reason why in a fuzzy system a particular case could not a good example of more than one class (have high degrees of belonging to more than one class). In the FCM method they will end up with nearly equal fuzzy memberships but small memberships compared with a case which is a good example of only one class.

### 2.2. Possibilistic c means

To address the shortcomings of the FCM introduced by the condition expressed in Equation 4, the Possibilistic $c$ means (PCM) was introduced by Krishnapuram and Keller (1993). Here the membership is solely dependent on the distance of an object from the prototype for that class, which is to say, Equation 4 is relaxed to be Equation 5:

$$\sum_{i=1}^{c} \mu_{ij} \leq 1 \text{ for all } j. \quad \text{Equation 5}$$

Here the possibility of object $i$ being in class $j$ is given by

$$\mu_{poss(ij)} = \frac{1}{1 + \left(\frac{d_{ij}}{\eta_i}\right)^{\frac{1}{m-1}}} \quad \text{Equation 6}$$

where $\mu_{poss(ic)}$ is the possibility (also sometimes known as the fuzzy membership) of an object $i$ in the set or class $c$, and
$\eta_i$ describes the shape of the distribution (Krishnapuram and Keller, 1993).

These equations (1 and 6) can be easily implemented for the OAC. For every OA classified, the distance to the centre of every candidate cluster is recorded. These are recorded in tables as DIFF variables and are directly equivalent to $d_{ic}$ in equations 1 and 6.

## 3. First results

### 3.1. Fuzzy mapping

Figure 1 shows maps of the fuzzy memberships (after equation 1) for each of the seven Supergroups for much of Leicester City. It shows that in any one Supergroup relatively few OAs have large fuzzy memberships, and the fuzzy memberships themselves are not large (maximum value in any mapping 0.642 for Countryside). Many of the Supergroups are relatively compact, including City Living and Countryside, while others are made up of a number of compact areas, including Constrained-by-Circumstance, Blue Collar Communities, and Prosperous Suburbs. Multicultural Communities is relatively compact covering most of the north and east of the city, but with many OAs with larger memberships punctuating the area. Typical Traits is the class which is worst characterised by any particular set of variables. It presents a pattern with many small clusters, and few large memberships in the countryside or city centre. Maximum and minimum fuzzy memberships in each of the seven Supergroups values are listed in Table 1. All Supergroups can be seen to have unimodal distributions strongly negatively skewed towards very small memberships.

### 3.2. Possibilistic mapping

In contrast Figure 3 shows the result of mapping the possibilities of class membership. These have a much larger range of values with the maximum in all mappings being over 0.85 and Countryside going as high as 0.94. The mappings are a major contrast to those for fuzzy memberships. City Living (perhaps the most compact Supergroup in the fuzzy memberships) is seen to be extended over a large part of the city, with a more punctuated habit with many local large values. Countryside is as compact as in the fuzzy mapping, but with a greater area showing darker, giving an impression of a tighter ring around the city. The multiple compact areas in Constrained-by-Circumstance and Blue Collar Communities are expanded both forming almost a semi-ring to the west of the city. Prosperous Suburbs is very different from the fuzzy mapping forming a much more continuous area outside the city (but especially to the west). Here Multicultural Communities is seen to spread across the whole city with large values almost everywhere except in the city centre and in the surrounding countryside. Typical Traits can be seen to be a very general class geographically, with large possibility values in the countryside as much as in the city, although again the city centre itself has small values. Maxima and minima for the possibilistic $c$ means memberships for each Supergroup are presented in Table 2. These clearly show a variety of patterns of distributions. Some seem almost normally distributed (Typical Traits), some are multimodal (Prosperous Suburbs) while others are negatively skewed (Blue Collar Communities). Numerical ranges are very different from those of the fuzzy memberships (Figure 2), being consistently larger, and only rarely dropping below 0.3 membership.

## 4. Further work

Novel ways of looking at the OAC results together with interesting visualisations can be derived from this analysis. Indeed alternative class assignments for Output Areas can be achieved. These will be reviewed in the presentation.

## 5. References

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, NewYork.

Bezdek, J. C., Ehrlich, R., & Full,W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences* 10 (2-3), 191−203.

Burrough, P.A., and McDonell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press

Burrough, P.A., van Gaans, P.F.M., and MacMillan, R.A., 2000. High-resolution landform classification using fuzzy *k*-means. *Fuzzy Sets and Systems* 113 (1), 37-52.

Harris, R., Sleight, P. and Webber, R., 2005. *Geodemographics, GIS and Neighbourhood Targeting*. Wiley, Chichester.

Krishnapuram, R., and Keller, J.M., 1993. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems* 1 (2), 98-110.

Vickers, D.W., and Rees, P.H., 2007. Creating the National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society*, Series A, 170 (2), 379-403.

Vickers, D.W., and Rees, P.H., 2006. Introducing the National Classification of Census Output Areas, *Population Trends*, 125.

**Table 1 -** Fuzzy Membership distribution properties of the Supergroups

| | Minimum | Maximum |
|---|---|---|
| Blue Collar Communities | 0.0559 | 0.4905 |
| City Living | 0.0329 | 0.4131 |
| Countryside | 0.0473 | 0.6418 |
| Prosperous Suburbs | 0.0462 | 0.5602 |
| Constrained by Circumstances | 0.0365 | 0.3903 |
| Typical Traits | 0.0866 | 0.5119 |
| Multicultural Communities | 0.0284 | 0.5102 |

**Table 2 -** Possibilistic Membership distribution properties of the Supergroups

| | Minimum | Maximum |
|---|---|---|
| Blue Collar Communities | 0.2561 | 0.9189 |
| City Living | 0.3060 | 0.8560 |
| Countryside | 0.2324 | 0.9391 |
| Prosperous Suburbs | 0.2283 | 0.9005 |
| Constrained by Circumstances | 0.2905 | 0.8849 |
| Typical Traits | 0.2832 | 0.9359 |
| Multicultural Communities | 0.2561 | 0.8768 |

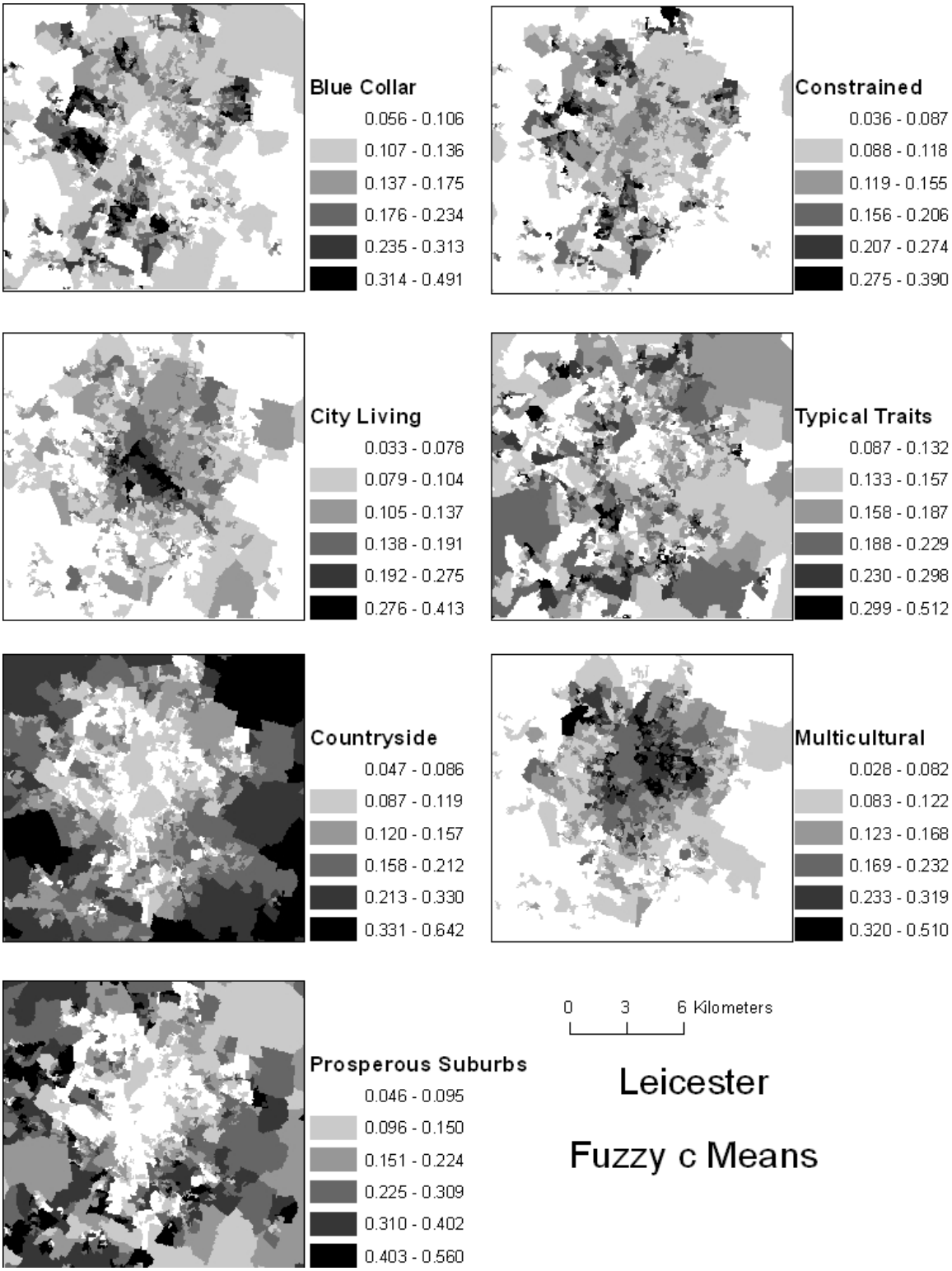



**Figure 1** – Fuzzy *c* Means mappings of each of the seven Supergroups in Leicester

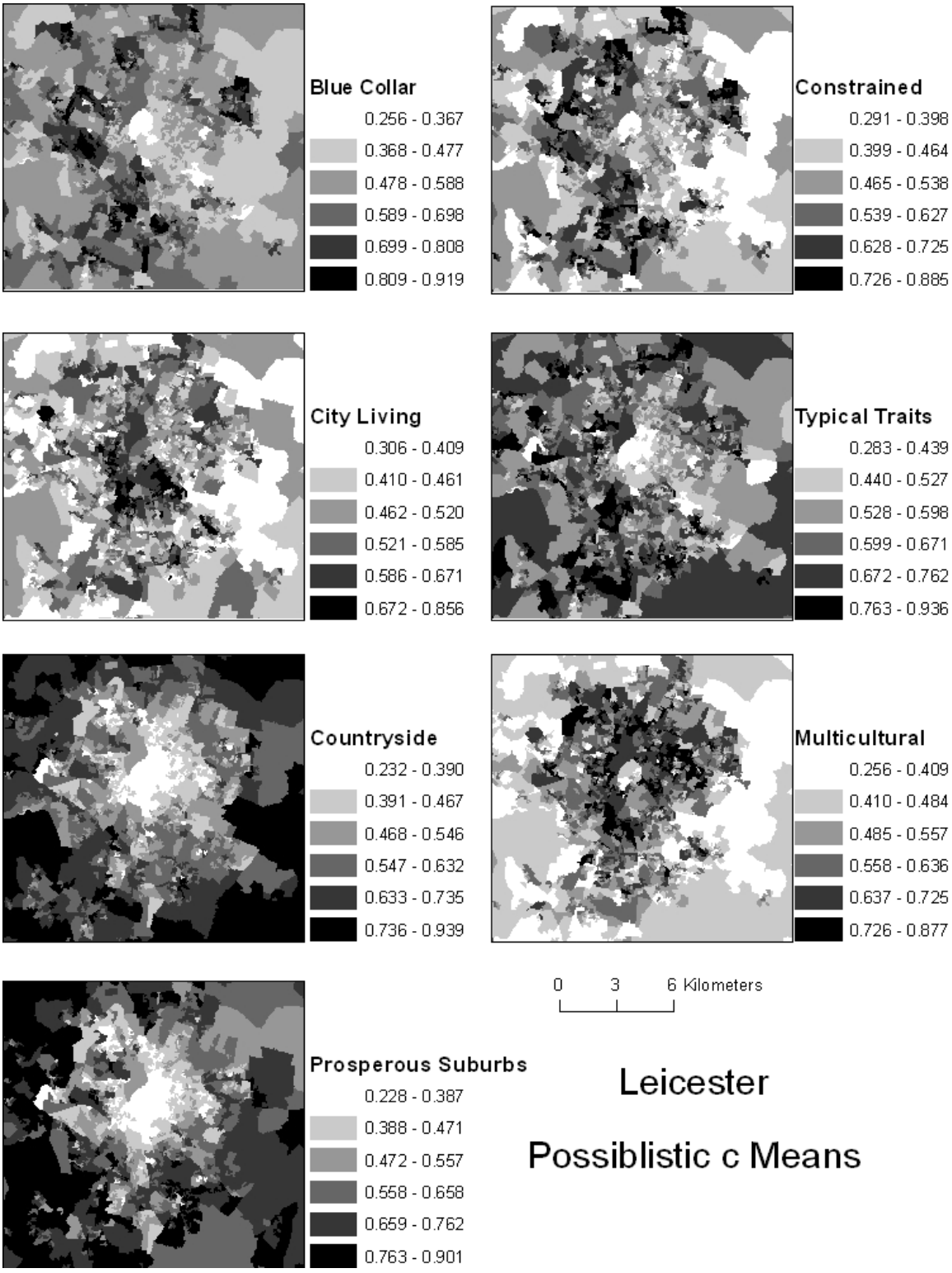



**Figure 2** – Possibilistic *c* Means mappings of each of the seven Supergroups in Leicester

# OAC Explorer: Interactive exploration and comparison of multivariate socioeconomic population characteristics

Aidan Slingsby[1], Jason Dykes[1], Jo Wood[1] and Robert Radburn[2]

[1]giCentre, Department of Information Science,
City University London, Northampton Square, London, EC1V 0HB, UK
Tel. +44 (0)20 7040 0180
{sbbb717 | jwo | jad7}@soi.city.ac.uk, http://gicentre.org/

[2]Leicestershire County Council
County Hall, Glenfield, Leicestershire, LE3 8RJ, UK
Robert.Radburn@leics.gov.uk

KEYWORDS: visualization, geodemographics, demography, oac, census

## 1. Introduction

National censuses are valuable sources of population data, but are too detailed their raw forms for most end-users. Geodemographic classifiers are derived data products that simplify these data by using statistical clustering to characterise local population-profiles in small areas. They are used widely in both the public and private sectors for population studies, sampling, marketing and service provision. OAC (Vickers and Rees, 2007) is an example of a geodemographic classifier that is based on the statistical clustering of 41 census variables (Table 2), categorising population by Output Area (OA; ~125 households; >200,000 in UK) into 7 super-groups and 21 groups[1] (Table 1). Considerable generalisation is involved as 7, 21 or 52[1] classes are produced from 41 continuous variables. Assigning OAs to single classes conceals the fact that most OAs are atypical cases sharing characteristics of many classes.

Local authorities are increasingly encouraged to use OAC to improve services (DCLG, 2009, p73). Leicestershire County Council (LCC) uses OAC for population profiling at different spatial scales (e.g. OAs, wards, postcode districts, postcode sectors) for service provision planning and for linking health, crime and other datasets.

Interactive visualization techniques for exploratory data analysis allow real-time selection and data filtering on demand in response to changing lines of enquiry. This "overview first, zoom, filter then details on demand" approach (Shneiderman, 1996) is the basis of many tools and environments (e.g. Dykes, 1998). LCC are interested in quantifying uncertainty in OAC classification, as this has implications for its utility, through visualization. OAC Explorer is designed to address the types of questions asked by LCC through a fast interactive visual interface epitomising this approach:

- Access to the census variables values used in the creation of OAC for all >200,000 OAs.
- Depicting uncertainty in the classification.
- Comparison with the typical cases of super-groups and groups.
- Comparison across different spatial scales.

[1] OAC's further 52 subgroups are not considered here.

**Table 1.** OAC super-groups and groups.

| Super-group | Group |
|---|---|
| Blue Collar Communities | Terraced Blue Collar |
| | Younger Blue Collar |
| | Older Blue Collar |
| City Living | Transient Communities |
| | Settled in the City |
| Countryside | Village Life |
| | Agricultural |
| | Accessible Countryside |
| Prospering Suburbs | Prospering Younger Families |
| | Prospering Older Families |
| | Prospering Semis |
| | Thriving Suburbs |
| Constrained by Circumstances | Senior Communities |
| | Older Workers |
| | Public Housing |
| Typical Traits | Settled Households |
| | Least Divergent |
| | Young Families in Terraced Homes |
| | Aspiring Households |
| Multicultural | Asian Communities |
| | Afro-Caribbean Communities |

## 2. Data

OAC is an open classifier whose creation methodology and census variables (Table 2) are freely available (SASI, 2009). Fuzzy membership is provided (SASI, 2009) as distances (similarity measures) from each of the super-group cluster centroids of each super-group for each OA. We normalise these distances to the maximum for each OA, inverting these to produce a measure of super-group ($g$) typicality ($T_g$):

$$T_g = 1 - \frac{d_g}{\max(d)} \tag{1}$$

*($d_g$ = distance to cluster g, max(d) = distance to the furthest cluster and $d_g$=min(d) when closest cluster is used)*

Scaled entropy ($E$; Fisher *et al*, 2004) is a measure of classification reliability, scaled between 0 (unreliable – where the OA is as similar to each super-group) and 1 (reliable, where the OA is typical of just one super-group):

$$E = \frac{-\sum_{i=7}^{i=1} d_i \times \log(d_i)}{-\log(1/7)} \tag{2}$$

*($d_i$ =distance to super-group i)*

**Table 2.** OAC's 41 census variables.

| | |
|---|---|
| Age | 0-4 **[v1]**, 5-14 **[v2]**, 25-44 **[v3]**, 45-64 **[v4]**, over 65 **[v5]** |
| Ethnicity | Indian/Pakistani/Bangladeshi **[v6]**, Black **[v7]** |
| Country of birth | Born outside UK **[v8]** |
| Population | Population density **[v9]** |
| Household | Separated/divorced **[v10]** |
| Household composition | Single non-pensioner **[v11]**, single pensioner **[v12]**, lone parent household **[v13]**, two adults no children **[v14]**, with non-dependents **[v15]** |
| Housing tenure | Public rent **[v16]**, private rent **[v17]** |
| Housing type | Terraced **[v18]**, detached housing **[v19]**, flats **[v20]** |
| Housing quality | No central heating **[v21]**, rooms per household **[v22]**, people per room **[v23]** |
| Education | HE qualification **[v24]** |
| Socio-economic | routine occupation **[v25]** |
| Travel | 2+ car household **[v26]**, public transport to work **[v27]**, work from home **[v28]** |
| Health and care | Limiting Long Term Illness **[v29]**, provide unpaid care **[v30]** |
| Employment | Full-time students **[v31]**, unemployed **[v32]**, working part-time **[v33]**, economically inactive looking after family **[v34]** |
| Industrial sector | Agriculture/fishing **[v35]**, mining/quarrying/construction **[v36]**, manufacturing **[v37]**, hotel/catering **[v38]**, health/social work **[v39]**, financial intermediation **[v40]**, wholesale/retail trade employment **[v41]** |

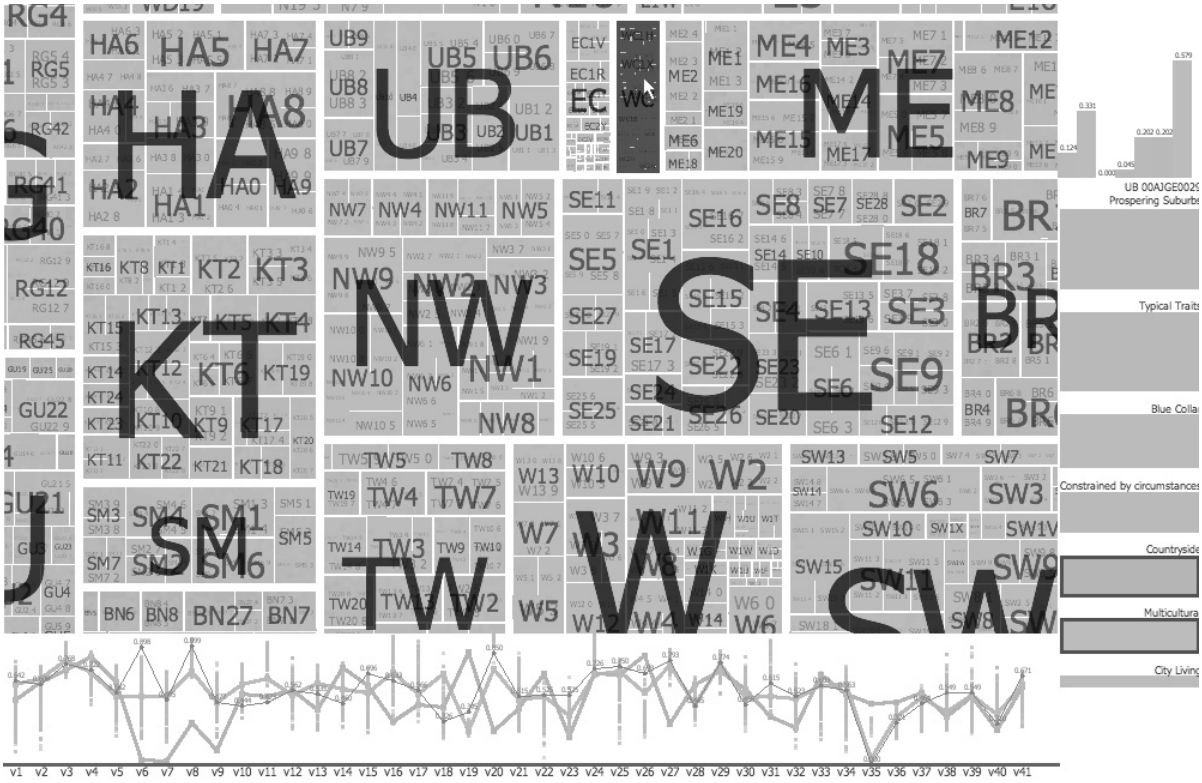



**Figure 1.** OAs in London, sized by population and coloured by super-group. Variable values displayed on parallel axes (below). The degree of typicality (*T*; equation 1) of an OA (indicated by the mouse pointer) to each super-group is shown in the barchart (top right).

## 3. Visualization techniques

Our new colour[2] scheme for OAC (Wood *et al*, 2010) uses 7 perceptually uniform hues for each super-group at constant lightness. In Table 1, we vary hue slightly for each group. It is not possible to

[2] Colour PDF version of this paper is available from http://gicentre.org/papers/slingsby_oacexplorer_2010.pdf

produce 21 distinguishable hues, so these colours are designed to reflect heterogeneity within super-groups rather than for class lookup. We have computed palettes for each of the 21 hues whereby lightness varies perceptually-linearly and is comparable across hues and use this in Figure 2 where lightness indicates typicality (*T*; equation 1; left) and entropy (*E*; equation 2; right).

Figure 1 contains a screenshot for the London area. The hierarchical rectangular cartogram (spatial treemap; Wood and Dykes, 2008) sizes OAs by their population within the postcode hierarchy. We use this representation over standard cartographic projections because it normalises by population density (population-dense areas have a greater map area so the detail is more easily resolvable), is space-filling (makes efficient use of space – important as there are >200,000 elements to map), retains a high degree of spatial structure (Slingsby *et al*, in press) and is already in use within LCC. The use of postcode geography places OAs in a widely-recognised spatial frame of reference allowing selection at different spatial levels – all OAs in the WC postcode area in this case (the strong blue colour indicates WC is selected). Each axis in the parallel coordinate display corresponds to a census variable (Table 2). Values for the OAs in WC are shown as dots with typical 'Multicultural' and 'Countryside' profiles as thicker lines (selected on the right). The thin red line in the parallel plot shows the census profile of the selected OA (indicated by the mouse pointer) and allows its relationship to other super-groups ('Countryside' in this case) to be seen. This interface allows national (typical cases for super-groups), regional (selected set of OAs; those in WC) and individual OAs to be selected and compared. Uncertainty in OA classification can be explored through colour lightness that indicates typicality (in Figure 2, left) and entropy (in Figure 2, right). The barchart for the selected OA shows proximity to all super-group centroids concurrently. Here, it is quite typical of 'Multicultural', but is also similar to other super-groups – particularly 'City Living' – and so has low entropy. This is interesting as many inner London OAs fall into these two classes. 'Typical Traits' show low entropies – the OAs of which form a tight cluster.

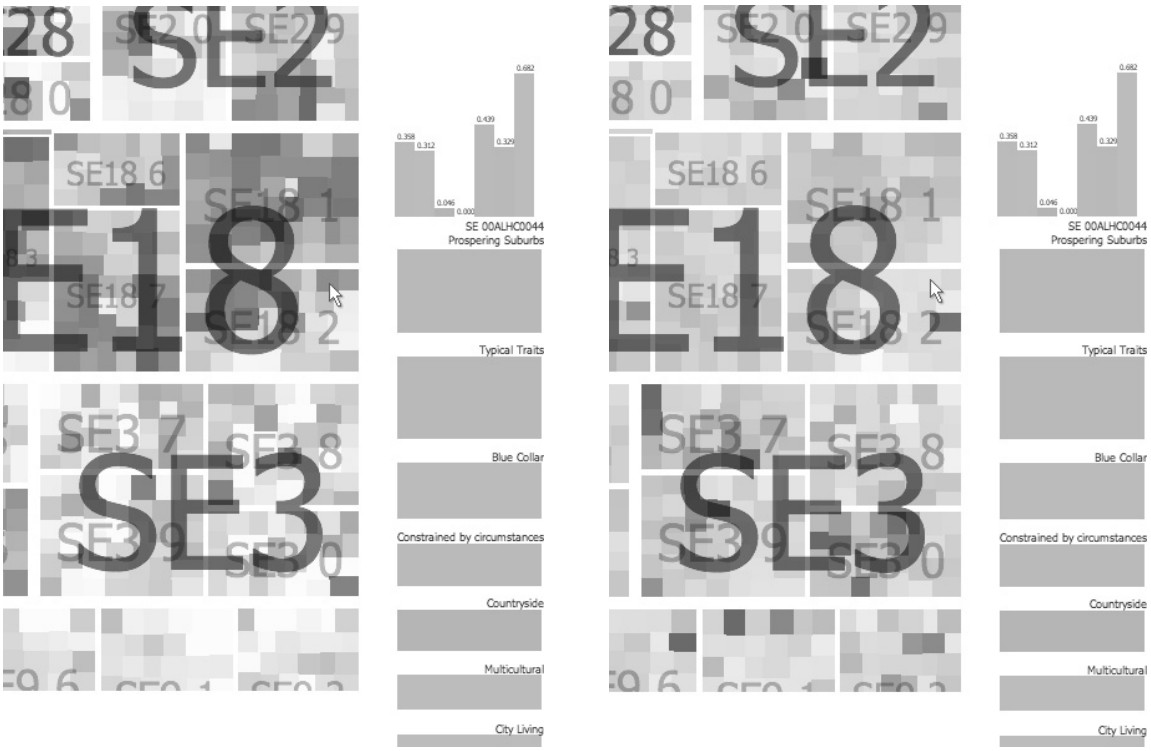



**Figure 2.** Screenshot excerpts that use *lightness* to show the typicality (*T*) of OAs to their closest super-group (left) and the scaled entropy (*E*) of classification (right).

Figure 3 shows the Leicester area. A 'Multicultural' core is distinguishable from the 'Prospering suburbs' and 'Countryside'. Box-plots of the 25$^{th}$, 50$^{th}$ and 75$^{th}$ percentiles indicate variation in the census variables for 'Multicultural' OAs in the visible screen area. The middle two quartiles of v7,

v20, v37 and v18 (Table 2) are significantly above the typical case for this super-group. Variables are sorted on the selected OA here enabling variables with atypical values to be identified.

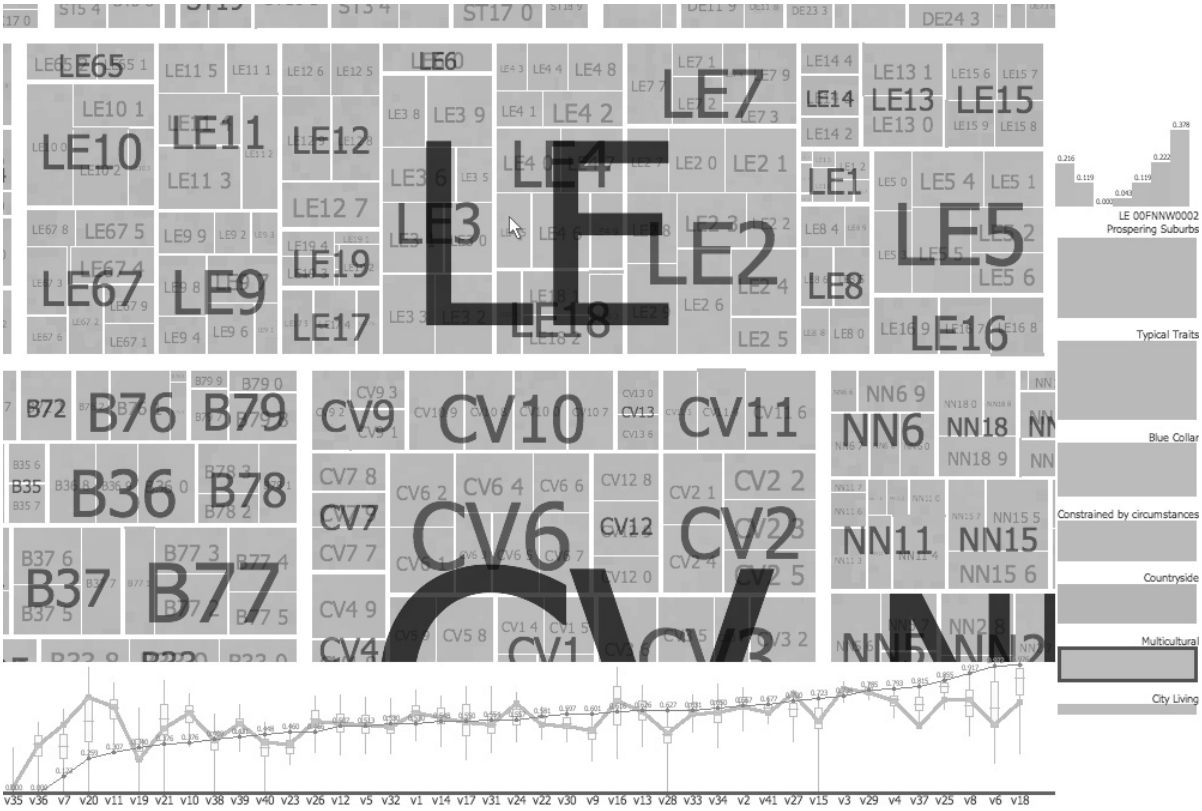



**Figure 3.** Screenshot for Leicester with Coventry to the south showing typical 'Multicultural' with box-plots summarising variation in the current field of view. The red line is the OA indicated with the mouse pointer, and variables are sorted by their magnitude on this line.

Figure 4 focuses on LE7. Variable values are shown as differences from the typical 'Prospering Suburbs' case. As expected, the 'Prospering Semis' subgroup deviates little from this (most notably, v19) but values for proportion of Indian/Pakistani/Bangladeshi (v6) vary significantly. This is very different to the national distribution (see Figure 5).

## 5. Conclusion and ongoing work

Geodeomographic classifiers help local authorities and others understand local populations through generalisation by categorising them into meaningful groups. We show that desktop computers are now powerful enough to provide rapid and interactive access to OAC alongside its original census variable values through which the effects of this generalization can be explored and evaluated. This visualization allows the reliability and uncertainty in classification to be determined at different spatial granularities and places. Ongoing work with practitioners in local government is developing and evaluating techniques for using visualization to meet their specific needs.

## 6. Acknowledgements

Thanks to Dan Vickers for providing the original OAC input variables. Aidan is funded by the Willis Research Network and Robert was supported by ESRC UPTAP User Fellowship RES-163-27-0017. The National Statistics Postcode Directory was obtained through UKBorders/Edina and other OAC-related data were from http://www.sasi.group.shef.ac.uk/area_classification/.

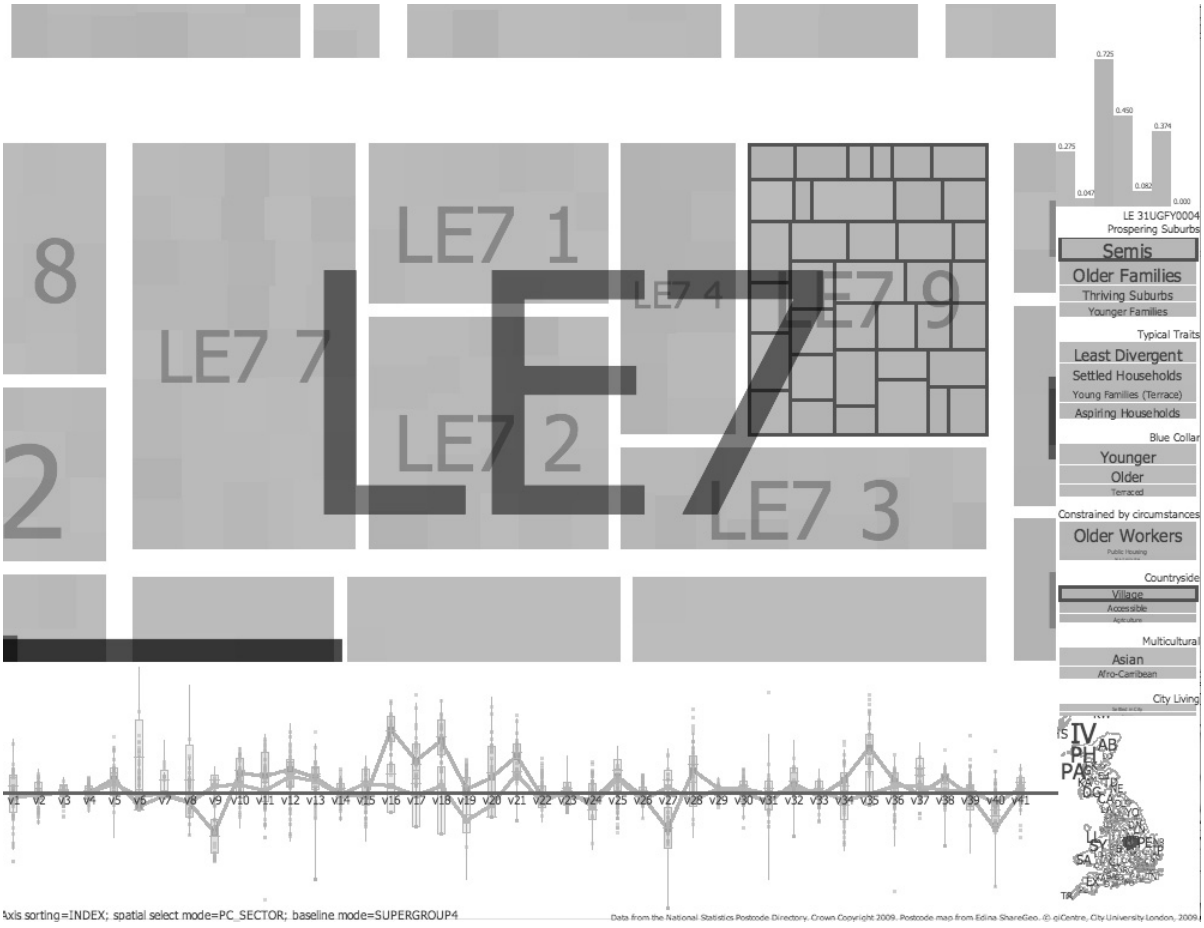



**Figure 4.** Screenshot focusing on LE7 with OAs of LE7 9 selected (its OAs are outlined in blue). The parallel plot shows deviation from 'Prospering Suburbs' super-group.

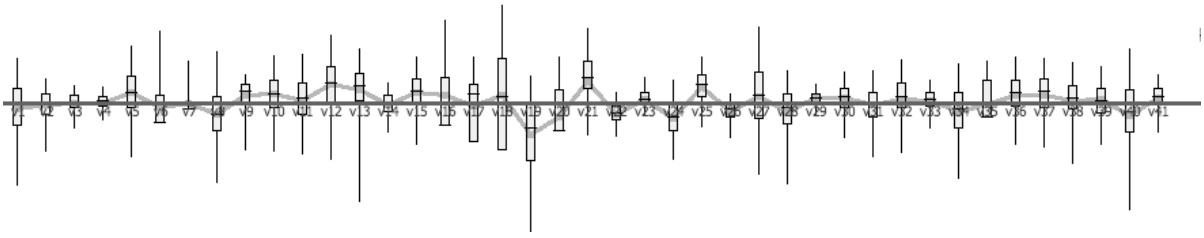


**Figure 5.** Box plot showing national variation in 'Prospering Semis'.

## References

DCLG (UK Department of Communities and Local Governement) (2009). Supporting local information and research: Understanding demand and improving capacity. *http://www.communities.gov.uk/publications/communities/supportinglocalresearch*

Dykes, J., (1998), Cartographic Visualization: Exploratory Spatial Data Analysis with Local Indicators of Spatial Association using TCL/TK and CDV, *The Statistician*, **47**(3), 485-497.

Fisher, P., Wood, J. & Cheng, T., (2004). Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. *Transactions of the Institute of British Geographers*, **29**, 106-128.

SASI (Social and Spatial Inequalities Group, University of Sheffield) (2009). *The National Classification of Census Output Areas*. *http://www.sasi.group.shef.ac.uk/area_classification/*

Shneiderman, B., (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*. IEEE Computer Society, p. 336 http://portal.acm.org/citation.cfm?id=834354

Slingsby, A, Dykes, J. & Wood, J., (in press), Hierarchical Rectangular Cartogram of OAC by GB Unit Postcode. *Journal of Maps*.

Vickers, D. & Rees, P., (2007). Introducing the National Classification of Census Output Areas. *Population Trends*, **125**, 380-403.

Wood, J. & Dykes, J., (2008). Spatially Ordered Treemaps. *IEEE Transactions on Visualization and Computer Graphics*, **14**(6), 1348-1355.

Wood, J, Slingsby, A. & Dykes, J., (2010), *Layout and Colour Transformations for Visualising OAC Data*, GIS Research UK 2010 *[this volume].*

**Biographies**

***Dr. Aidan Slingsby*** *is a Willis Research Fellow at the giCentre, City University London with research interests in designing, implementing and using geovisualization techniques for assessing data quality and variability and for visual data analysis.*

***Dr. Jo Wood*** *is a Reader in geographic information at the giCentre, City University London with research interests in geovisualization, terrain modelling and object oriented programming for spatial sciences.*

***Dr. Jason Dykes*** *is a Senior Lecturer at the giCentre, City University London undertaking applied and theoretical research in, around and between information visualization, interactive analytical cartography and human-centred design.*

***Robert Radburn*** *is Research and Intelligence Team Leader at Leicestershire County Council and held an ESRC UPTAP research fellowship at the giCentre, City University London to develop capacity for visual exploratory analysis in local government.*

# Parallel K-Means Clustering using Graphical Processing Units for the Geocomputation of Real-time Geodemographics

Adnan, M., Singleton, A.D., Longley, P.A.

University College London, Department of Geography, Gower Street, London, WC1E 6BT.
Tel: +44 (0)20 7679 0510 Fax: +44 (0)20 7679 0565
Email m.adnan@ucl.ac.uk, a.singleton@ucl.ac.uk, plongley@geog.ucl.ac.uk

KEYWORDS: Geodemographics, GIS, Clustering, CUDA, Geoweb 2.0

## 1. Introduction

Geodemographic classification categorise small geographic areas into a series of discrete categories that aim to represent the multidimensional characteristics of individuals living with these neighbourhoods. Real-time geodemographic classification is the vision for an online and automated web based system that enables users to build, visualise and test a bespoke classification within a short time period (probably in minutes). Within this context, the overarching aim of the paper is to evaluate some advances in parallel processing technology that may reduce the overall time taken to create a geodemographic.

Geodemographic representations are created using computationally intensive clustering algorithms which search the attribute space of a matrix of standardised input data comprising a row for each small area (however defined) and a column for each attribute measure. For example, Vickers and Rees (2007) used *k*-means clustering for the creation of the National Statistics Output Area Classification (OAC) with data derived from the 2001 Census of the Population. The *k*-means algorithm is a commonly used method for the geocomputation of geodemographic classification (Harris et al, 2005), however, in its original form, *k*-means is unstable and relatively sensitive to outlier values within the input data matrix. Because of this instability the algorithm requires multiple runs in order to ensure a robust result. For example, Longley and Singleton (2008) created a geodemographic classification using *k*-means with approximately 10,000 runs. However, this is a very time consuming process and problematic if geodemographic classifications are to be computed in near real time. For example, if the input data for the OAC (223061 rows and 41 columns) are clustered on a high specification computer where $k$ =52, then the processing time for this to converge is around 1.4 seconds. For this classification to be run 10,000 times, a user would have to wait for around 3 hours 53 minutes for their results.

Several improvements have been demonstrated in the *k*-means clustering algorithm aiming to enhance computational efficiency. For example, Reynolds *et al* (2004) described "k-means++" which included a new way of choosing the initial seeds. This method selects these initial centres based on the density of data points and improved the overall processing time because clusters converge more quickly. However, even with these improvements, computational performance still remains a concern. This paper proposes a parallel implementation of *k*-means using NVIDIA's Computer Unified Device Architecture (CUDA)[i]. CUDA allows different processes to run in parallel on the Graphical Processing Units (GPUs) of NVIDIA's graphics cards.

## 2. Computer Unified Device Architecture (CUDA) and the Parallel Implementation of *k*-means

CUDA is a general-purpose parallel computing architecture that uses the GPUs of NVIDIA graphics cards to solve complex computational problems. A typical CUDA enabled NVIDIA graphics card has a number of GPUs and a set of memory capable of storing a reasonably large amount of data. For example, "GeForce 8400M GT" graphics card has 16 GPUs and 512MB of internal memory. CUDA requires that the computational problem to be programmed in the C language for parallel processing.

In order to improve the overall efficiency of *k*-means the remainder of this section reviews its implementation under CUDA. The *k*-means algorithm seeks to find a set of cluster centroids that minimises expression (1) below.

$$V = \sum_{j=1}^{k} \sum_{i=1}^{k} (x_j - \mu_i)^2 \qquad (1)$$

where $k$ is the number of clusters, and $\mu_i$ is the mean centroid of all the points $x_i$ in cluster $i$.

The *k*-means algorithm begins by randomly assigning a set of $k$ seeds within the data matrix and proceeds to allocate each data point to its nearest seed within the multidimensional space. A cluster centroid value is then calculated for each cluster, and a new partitioning of the data points is made around the new set of centroids. The centroids are then recalculated for the new clusters of points, and the algorithm repeats these steps until a convergence criterion is met (usually when the switching of data points no longer takes place between the clusters).

Some parallel implementations of *k*-means exist with CUDA. For example, Takizawa and Kobayashi (2006) proposed a parallel *k*-means solution for solving "image texture size problem". Hall and Hart (2004) also proposed another parallel solution for solving the problem of "limited instance counts and dimensionality" for complex shapes. However, these implementations only work in specified environments and there are no global parallel *k*-means solutions that are suitable for creating geodemographic classifications.

Our proposed parallel *k*-means algorithm via CUDA works as follows:

Total number of runs is specified by $N$.

a) Central Processing Unit (CPU) prepares the data points and counts the number of GPUs available on the NVIDIA graphics card. Afterwards the CPU uploads the data points and code instructing one *k*-means run to each GPU.

b) GPU performs *k*-means clustering on the data points by minimizing expression (1). When an optimal solution is achieved, GPU returns the result to CPU and claims the next *k*-means run from CPU if there are any.

c) CPU stores the results returned by GPUs in a local data structure contained in Random Access Memory (RAM). CPU keeps on delegating requests to GPUs until number of runs are less than $N$.

d) If number of runs is equal to $N$, CPU compares the "within sum of squares distance" optimisation criteria of all the runs.

e) The optimal solution is the one that has minimum "within sum of squares distance".

## 3. Results

The remainder of this paper demonstrates the implementation of the algorithm outlined in the previous section. The hardware used for this evaluation comprised an "Intel Core2 Duo 2.10GHz" CPU, 4GB RAM, and "GeForce 8600M GS" NVIDIA graphics card. The graphics card has 16 GPUs and 512 MB of RAM. For comparison we used the input data for the National Statistics Output Area Classification (Vickers & Rees, 2007) aggregated at three geographical levels: Output Area (OA); Lower Super Output Area (LSOA); and Ward for London. The following preliminary results demonstrate a comparison of the "computational time" in section (3.1) and the "time throughput" increased by using "Parallel *k*-means" in section (3.2). "Time Throughput" demonstrates the gain in computational time when running "Parallel *k*-means" instead of *k*-means clustering.

### 3.1 Comparing "Computational Time"

In order to compare "computational time" we ran *k*-means and parallel *k*-means for (*k*=2-30) cluster solutions at all of the three different geographic levels using the London datasets, and then compared the time taken for each algorithm to converge on a specified number of clusters. For each value of *k*, each algorithm was run 100 times and the results are shown in Figures 1-3. "Computational time" represents the time an algorithm takes to complete 100 iterations for each value of k.

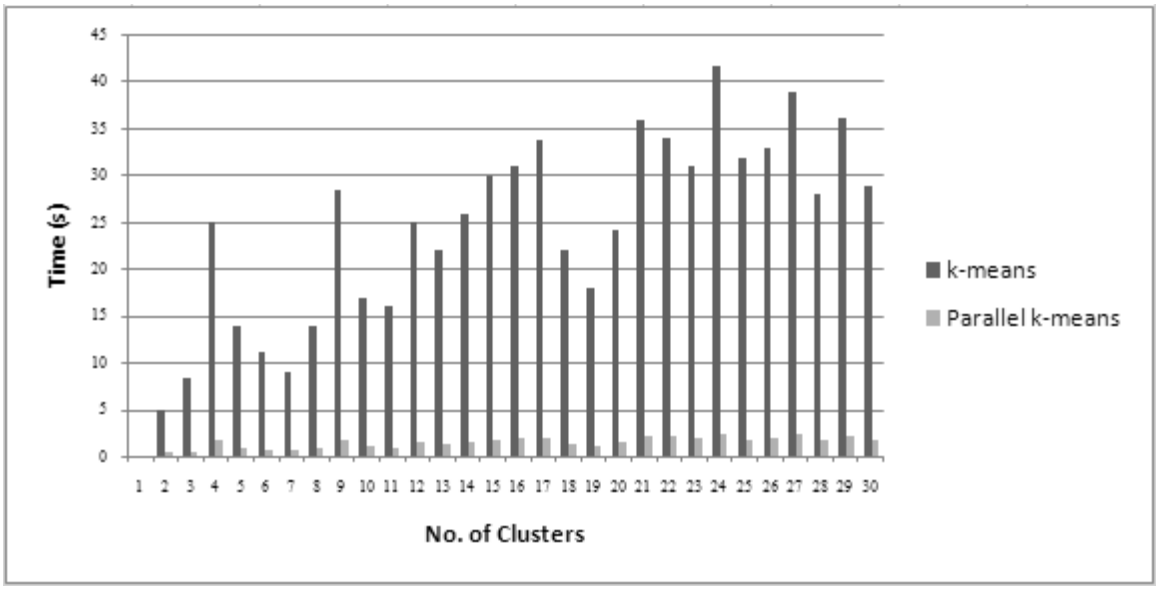



**Figure 1:** Output Area (OA) level results for the two clustering algorithms

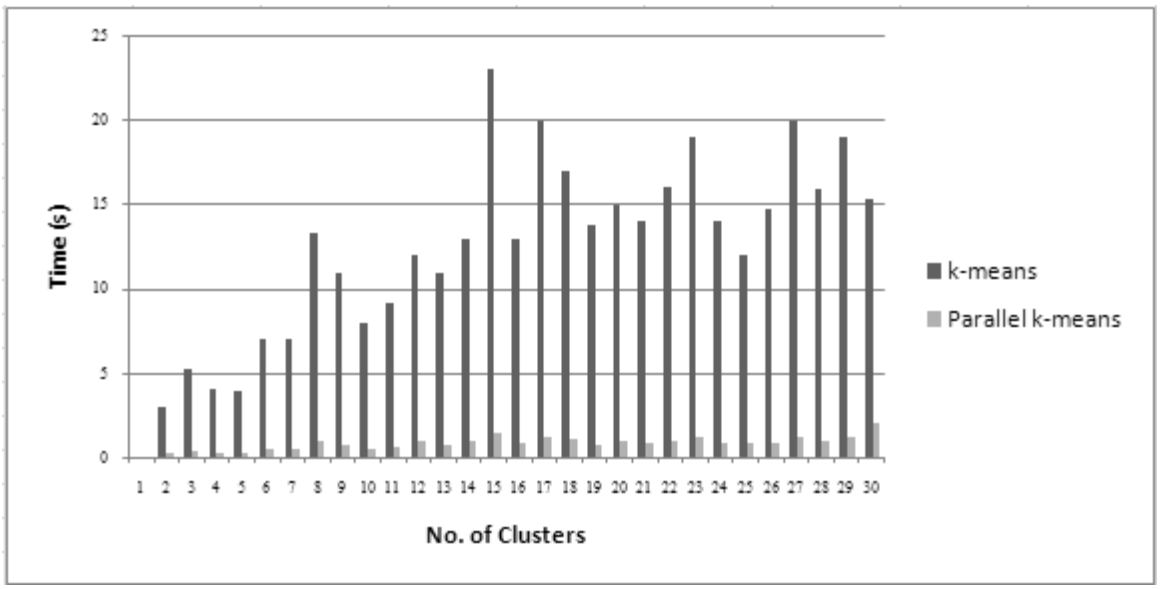



**Figure 2:** Lower Super Output Area (LSOA) level results for the two clustering algorithms

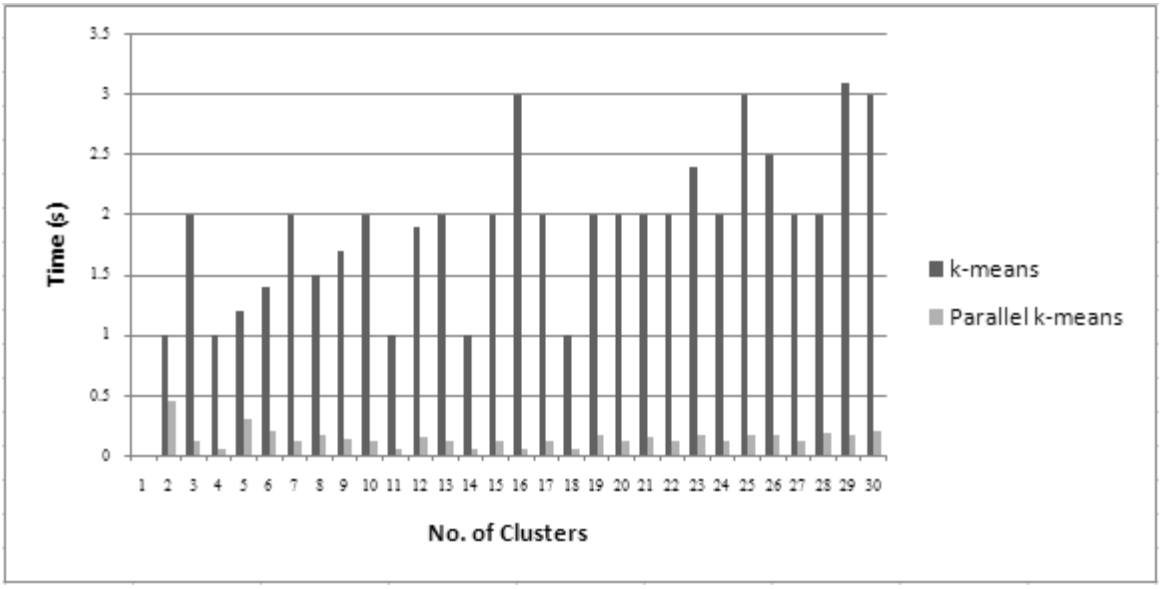



**Figure 3:** Ward level results for the two clustering algorithms

These results indicate that the parallel implementation of *k*-means out performs *k*-means considerably in terms of computational time across all different spatial levels featured in the evaluation.

#### 4.2 "Time Throughput" increased for "Parallel k-means"

Figures 4 - 6 show the relationship between "time throughput" achieved using "Parallel *k*-means" and the number of clusters (*k*=2-30) by using *k*-means and parallel *k*-means again at all of the three different geographical levels in the data covering London. For each value of *k*, each algorithm was run 100 times and the results are shown in Figures 4-5.

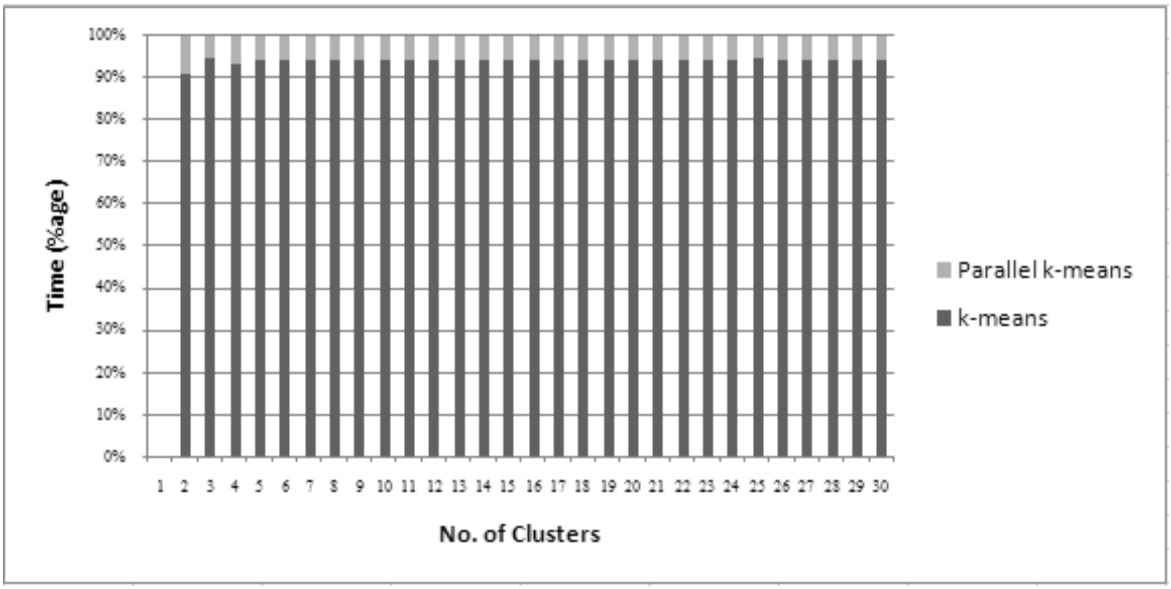



**Figure 4:** Output Area (OA) level results for the two clustering algorithms

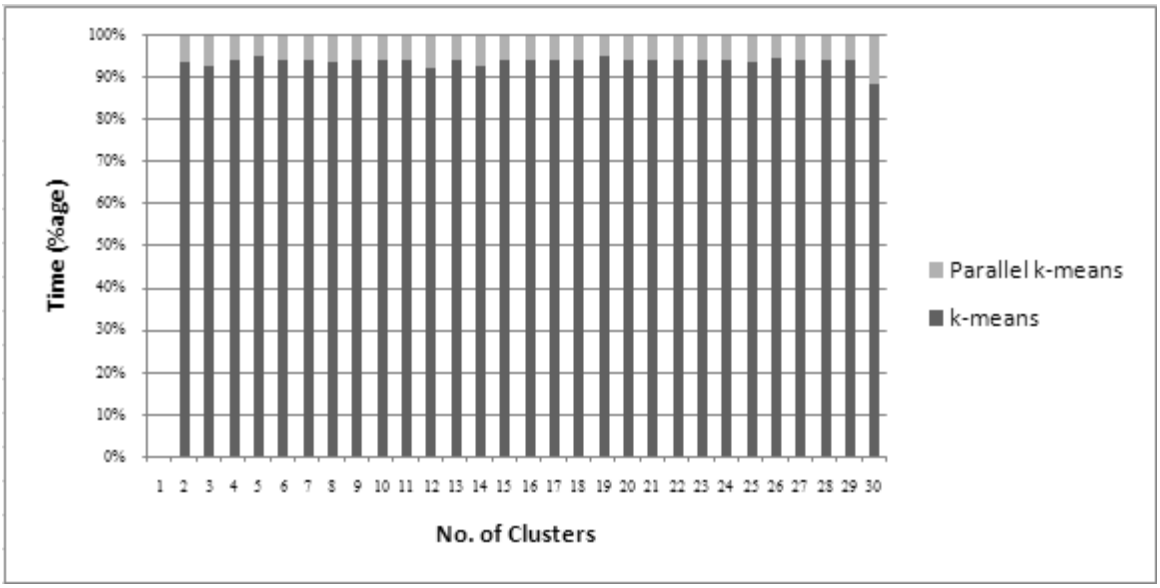



**Figure 5:** Lower Super Output Area (LSOA) level results for the two clustering algorithms

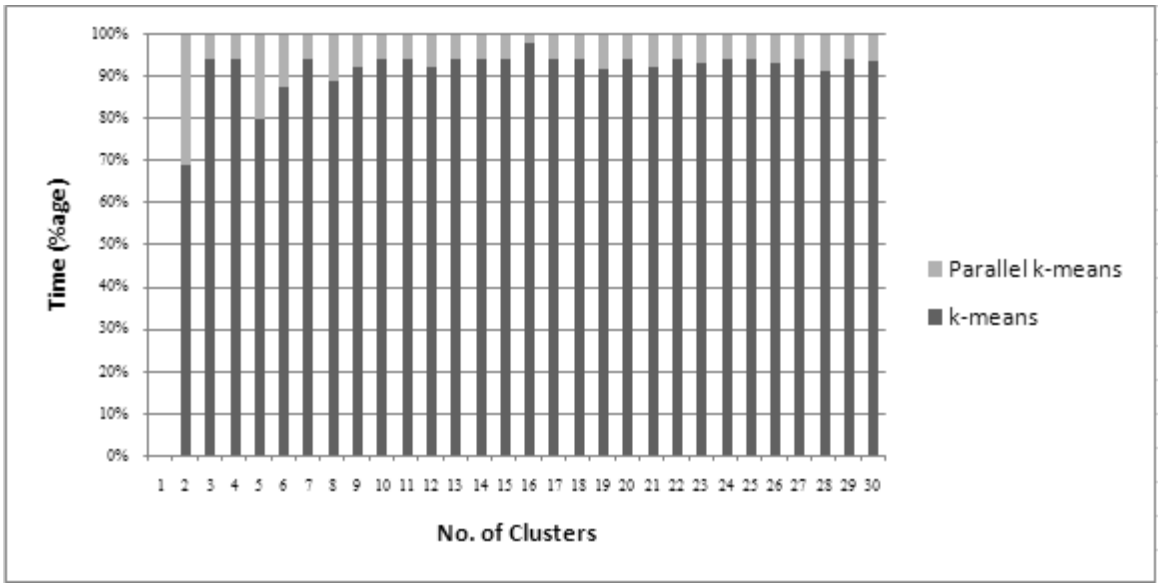



**Figure 6:** Ward level results for the two clustering algorithms

The above results demonstrate that for all geographical levels parallel *k*-means gives approximately 90% more time efficiency than *k*-means.

## 5. Conclusion and Future Research

This paper contrasted traditional *k*-means with the proposed "Parallel *k*-means" clustering algorithm for creating real time geodemographic classifications. It was found that the parallel implementation of *k*-means dramatically reduced the overall computational time for the clustering algorithm. This paper gives a proof of concept of using CUDA for the creation of real time geodemographic system. However, further work is required on more high powered NVIDIA graphics cards with more GPUs and memory. For example, the NVIDIA S2070 server contains 4 Telsa GPU which house 24gig memory and around 2000 cores.

## References

Hall, J.D., Hart, J.C. (2004). GPU acceleration of iterative clustering.

Harris, R., Sleight, P., Webber, R. (2005). Geodemographics, GIS and Neighbourhood Targeting. Wiey, London.

Office of National Statistics (2008). Ness data Exchange. From http://www.neighbourhood.statistics.gov.uk/HTMLDocs/downloads/NeSS-Data-Exchange-Technical-Implementation-Guide-v1.0.pdf.

Reynolds, A.P., Richards, G., Rayward-Smith, V.J. (2004) The Application of K-Medoids and PAM to the Clustering of Rules. Lecture Notes in Computer Science. 3177/2004, 173-178.

Singleton, A.D., Longley, P.A (2008). Creating open source geodemographic classifications for Higher Education applications.

Takizawa, H., Kobayashi, H. (2006). Hierarchical parallel processing of large scale data clustering on a pc cluster with GPU co-processing. J. Supercomput.,36(3):219–234, 2006.

Vickers, D.W. and Rees, P.H. (2007). Creating the National Statistics 2001 Output Area Classification. Journal of the Royal Statistical Society, Series A. 170(2), 379-403.

## Biographies

*Muhammad Adnan is working as a Knowledge Transfer Partnership Associate at Local Futures Ltd., and is also a second year PhD student. His research interests concern the automation of spatial data infrastructures, and the optimization of clustering algorithms to develop real time geodemographic classifications for public service delivery and decision making.*

*Dr Alex D Singleton completed his PhD in November 2007 and is now a Research Fellow in the Department of Geography and Centre for Advanced Spatial Analysis at University College London. His research interests include the ways in which geodemographic methods can be refined through modern scientific approaches to data mining, geographic information science and quantitative human geography.*

*Professor Paul Longley, Department of Geography, UCL. Paul Longley's research interests concerns the use of GIS and quantitative methods in urban analysis. He is the outgoing Editor-in-Chief of Computers, Environment and Urban Systems and reviews editor of Environment and Planning B as well as Deputy Director of Centre for Advanced Spatial Analyses at UCL.*

# DISAGGREGATING THE NIGERIAN POSTCODE: A STEP TO CREATING AN ENVIRONMENT FOR GEOMARKETING IN NIGERIA

Nicholas Babatope Allo

Centre for GIS, School of Geography, Geology and the Environment, Kingston University, Penrhyn Road, Kingston upon Thames, Surrey KT1 2EE
Email: K0740994@kingston.ac.uk

KEYWORDS: geomarketing, geodemographics, postcodes, basic spatial unit, spatial resolution

## 1. INTRODUCTION

The aggregation of small area datasets, into larger areas for representing and interpreting areal phenomenon, is suggested as the most appropriate approach for socioeconomic analysis by Harris and Longley (2002); they state, "it is easier to aggregate up from finer scales to identify broader geographical patterns than it is to try to reverse engineer, or 'deaggregate', coarse aggregate datasets".

Coupled with this, the evolution of Output Areas (OA) from pre-existing Enumeration Districts (ED), as found in Britain, is considered indicative of a drive to providing a granular view to broad socioeconomic information, and its assessment (Exeter et al., 2005; Bates, 2008; Vickers and Rees, 2006 and 2007). This has been made possible by using unit postcodes as the Basic Spatial Unit [BSU] for the collection of socioeconomic data.

## 2. POSTCODES AND THEIR RESOLUTIONS

Martin (1991) describes the postcode as, "the smallest indivisible part (e.g. a single terrace of houses) of a postman's 'walk' and contains an average of fourteen addresses"; Raper et al. (1992) provides five descriptive typologies, used in identifying locations associated to postcodes as: 1. Nominal, 2. Partially sequenced, 3. Topological, 4. Local geometry and 5. Global geometry and further describes postcodes as, 'a way of converting from a location related description (from a spatial format), into a useable contextual format'. Both the studies by Martin and Raper et al. can be described as advances on recommendations given in the Chorley Report (DOE, 1987) on using geographic information in the UK and its advocacy for postcodes.

In general, postcodes are a hierarchical structure to addressing and are used in many parts of the world in steering mail from an origin to a final destination, and their structures differ from country to country; in their simplest form, postcodes are described as 'labels that define a set of mail delivery points' (Raper et al., 1992). In Nigeria, postcodes were introduced during the later part of the year 2000 and these were intended to equally function as address labels and also, provide hierarchical structures to postal addressing in Nigeria. However, postcodes in Nigeria have not gained ground nor are they used despite all efforts to introduce them into the Nigerian postal system. A likely cause of failure, for postcodes in Nigeria, may be attributed to the spatial resolution applied in creating unit postcodes for postal services.

A BSU or unit postcode is expected to effectively, identify a set of individual addresses grouped together and identified by a given reference; also, unit postcodes should have an average of fourteen addresses, depending on the number of delivery points per postcode. However, in Nigeria the converse is the case as unit postcodes in Nigeria do not provide a granular resolution; instead, many streets are grouped together as unit postcodes. This is presented as a BSU or unit postcode within Nigeria (see Figure 1).

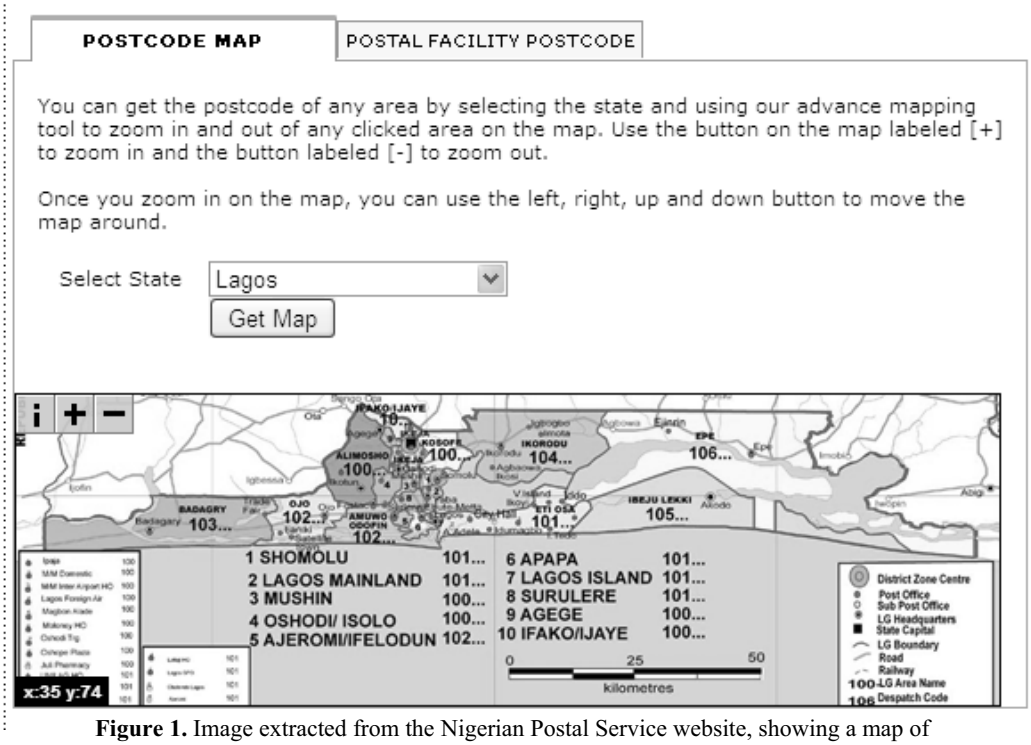



**Figure 1.** Image extracted from the Nigerian Postal Service website, showing a map of Lagos State which is part of a postcode zone and is further split into 10 postcode areas. Source: NIPOST, http://www.nipost.gov.ng/Default.aspx

The unit postcode in Nigeria does not offer any means of individually identifying an address; hence, despite introduced postcodes and suggested addressing formats (see Figure 2), topological descriptions for an address label are still predominant. On closer examination, postcode structures in Nigeria adopt spatial resolutions representing administrative boundaries, i.e. 'local government = postcode area' and 'an aggregation of states = postcode zones'.

The result of the above described is that aggregative properties of unit postcodes, enabling their use in the collection and management of socioeconomic data at an individual level, is not achievable with the existing spatial resolutions in use for the Nigerian postcodes. This deprives Nigerian policy makers of a capacity, to effectively identify underlying socioeconomic trends, suppressed by existing large spatial resolutions currently used.

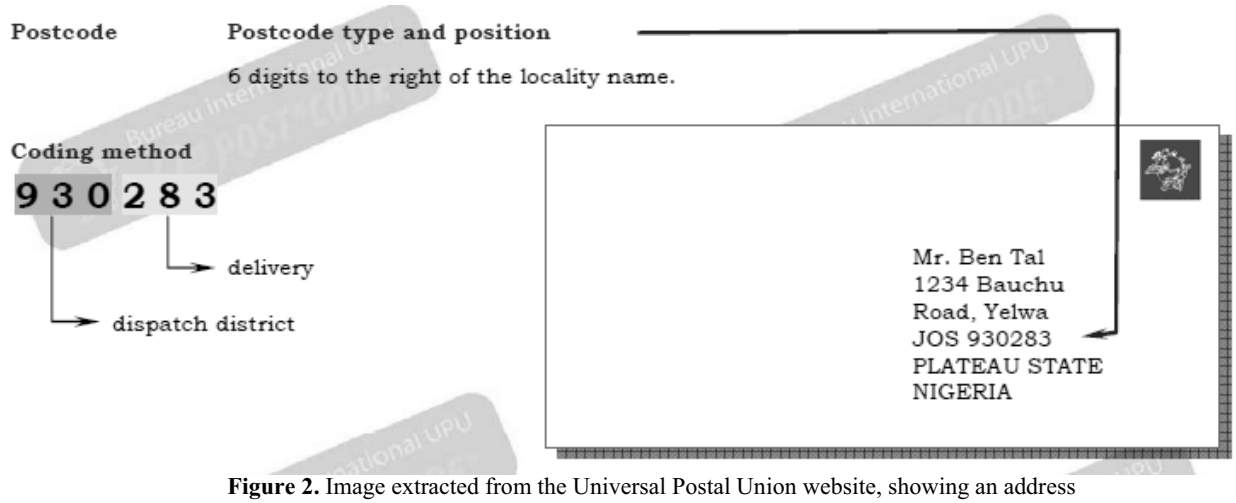



**Figure 2.** Image extracted from the Universal Postal Union website, showing an address label as intended for use with the postcode structure introduced by the Nigerian Postal service
Source: UPU, http://www.upu.int/post_code/en/countries/NGA.pdf

## 3. SOCIOECONOMIC ASSESSMENTS AND POSTCODES

Martin (1991 & 1992) reviews the application of postcodes with respect to handling socioeconomic data (see Figure 3); examined therein, are the approaches of associating grid references from postcodes to EA boundaries [using centroids] and also the challenges associated with the process of linking postcodes and census geographies while also handling census information.

In general, the above linkage is applied by Vickers and Rees (2006 and 2007) and used in creating the 2001 UK OA; this is further used in the creation of geodemographic profiles and in the application of geomarketing. The use of postcodes has also enabled the assessment of deprivation at an individual level with Samples of Anonymised Records [SAR]; it has further allowed the review of what is described as an Index of Multiple Deprivation [IMD] (Harris and Longley, 2002; Longley and Harris, 1999; DETR, 2000 and Fieldhouse and Tye, 1996).

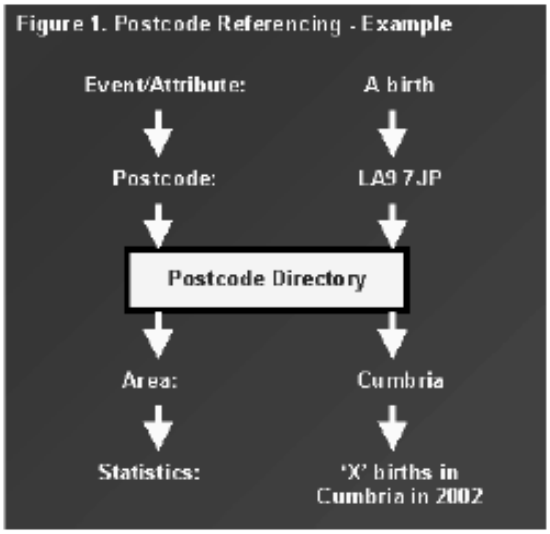



**Figure 3.** Image extracted from the Office of National Statistics website, showing the linkage between unit postcodes and the collection of socio-economic data.
Source: ONS, http://www.statistics.gov.uk/geography/geog_ref.asp

However, when considering the management and collection of socioeconomic information in Nigeria, along side the practise of collecting such information with a postcode or BSU, past studies identify attempts at mapping the distribution of poverty or deprivation (Coulombe and Wodon, 2007; Oladeji

and Abiola, 2000) used rather large geographies, such as district areas [equivalent to ward areas] in the collation and management of population data; this makes individual assessment of phenomena impossible. This may be attributed to earlier reasons outlined in the study by Prothero (1963). Prothero discusses likely issues affecting mapping of populations in Africa and lists some issues as: inadequate base maps available, incomplete or inaccurate population information and limited numbers of practising geographers required to undertake population mapping.

## 4. GEOMARKETING WITHIN A DEVELOPING COUNTRY

Geomarketing tools and methodologies [as well as geodemographics] have today become an integral part and an essential tool in the field of marketing; it is widely practised in developed countries and guides with regards a majority of marketing decisions made, such as customer segmentation, store site allocation, customer route analysis and customer targeting. Contrary to this, geomarketing as tool to enhancing market intelligence is yet to gain ground in developing countries, and marketing in places like Nigeria, is still based purely on records of customer spending; this greatly limits any marketing intelligence available or may be required which increases product and service sales and market penetration.

Another challenge identified in developing countries is the tendency for organisations to establish themselves in the commercial or state capital; in such cases, the general disposable income is reasonably higher than other regions within the country. A likely consequence of businesses located in a commercial or state capital, is a pricing mismatch of products and services against prospective customer segments.

This mismatch of product and service pricing, against customer prospects and the limited marketing intelligence, found in developing countries, are issues addressed by using geodemographics and geomarketing (Webber, 2004; Evans and Webber, 1994; Birkin, 1995; Evans, 1998; O'Malley et al., 1995 and 1997; Webber, 1985).

Key to implementing geomarketing in developing countries is having an information hierarchy that permits individual assessment, which in most cases is provided by using postcodes or a BSU (see section '2'); however, the current postcode structure in Nigeria does not provide an enabling environment for geomarketing. Hence, disaggregating the existing Nigerian postcode structure is seen as the first step to geomarketing in Nigeria.

Nigeria is believed would greatly benefit from the introduction of geomarketing, into both its public and commercial sectors. Within its commercial sector, businesses applying geomarketing would experience improved marketing intelligence and achieve pragmatic and effective location selection, using appropriate geodemographic information; it also reduces marketing dependence of businesses, on trending consumers spend for creating accurate customer segmentations or profiles. While in the public sector suppressed trends, resultant from using large spatial geographies for assessments, are better highlighted; meaning hidden pockets of phenomena [deprivation, health care, disability etc] are identified much faster, and are easier dealt with. This better enables the direction of policy, to meeting the needs of a wider populace.

## 5. DEVELOPING A NEW POSTCODE STRUCTURE FOR NIGERIA

In disaggregating the Nigerian postcode, the primary focus is reducing spatial resolution to a unit level from its existing administrative (local government area) level resolution. Using Kuipers (1978) TOUR model as guide to defining the spatial characteristics for postcode resolutions, we apply his descriptions of 'PATH' and 'PLACE' to define a cognitive extent; this compliments local knowledge of streets which also impacted creation of postcode in UK (ONS, 2005; Raper et al., 1992; Martin, 1991 & 1992).

From Kuipers TOUR model, 'PLACE' is described as inclusive of descriptions, of a local geometry of paths which intersect at a location; also, its local geometry further describes a relation between given or identified paths, their one dimensional orientation and radial headings; all within a local coordinate frame of intersection.

*PLACE*
NAME [name]
ON [list of PATHs]
STAR [local geometry data structure]

While a 'PATH', is described as a partial order of places, found along a path and represents, a partial knowledge possessed of place order, along such paths; it also possesses a one dimensional orientation with respect to directional order.

*PATH*
NAME [name]
ROW [partial order of data structure]

The new unit postcode resolution for Nigeria framed using Kuipers PLACE and PATH descriptors, used street topology (nodes & edges; lengths & intersections) by applying a semantic [named street] approach, to define a postcode units region of influence and applies dasymetric mapping principles [based on street topology and extracted land use types from satellite imagery] in defining an areal extent for derived postcode unit boundaries; hence, distribution of population within such an area is directly related to the network distribution and topology of streets (Boone, 2008; Mennis and Hultgren, 2006; Jiang, 2007; Jiang and Claramunt, 2004 and Porta et al., 2006). This offers an expected evenly distributed and individual level structure for unit postcodes in Nigeria.

Modifiable Areal Unit Problems (MAUP) were however considered when making some of these boundary extent decisions as some level of areal consistency is required for statistical information contained (Martin, 1991; Wong, 2009; Harris and Longley, 2004).

## 6. CONCLUSION

The procedure outlined aims to create postcode structures from the existing postcode hierarchy in Nigeria without an immediate consideration for grouping by dwelling types or characteristics as described by Raper et al. (1992) and Martin (1991).

This is as the intended process is not out to develop a complete postcode structure for Nigeria or a developing country, but to offer a framework of methodologies that may be applied irrespective of region; this would form the basis from which further development and definition to the postcode structure for a developing country would start, i.e. creating small area units suited to being described as postcodes.

Other classifications and descriptors to represent descriptions as postcodes would be better suited to organisations within these regions to apply; however, from the perspective of this research, the process of creating small area dataset for geomarketing has been initiated.

a

b

**Figure 4**. Images '**a**' and '**b**' are highlighted page sections, showing postcode areas found in Nigeria; while image '**c**'**,** showing the same postcode region, is reformatted from its original internet version obtained from the Nigerian Postal Service website. It highlights the inability to provide individual level mail routing required of unit postcodes and also reflects the non suitability of this spatial resolution as a BSU, for the collection of socioeconomic information at small area level
Source: NIPOST, 2003; NIPOST: http://www.nigeriapostcodes.com/views/

c

**Urban Area Postcode Lookup**

State: Lagos

Town: Lagos

Area: Shomolu Central

Street Name: Morocco Rd

Area: **Shomolu Central**
Postcode: **100231**
**Streets that Use the Postcode =>**

- A P. Petrol St. St.
- Abimbola St St.
- Abiodun St St.
- Abiola Lane St.
- Adebiyi St St.
- Adebowale St St.
- Adekoga St St.
- Adelawa St St.
- Aduroshakin St St.
- Agunbide St. St.
- Aiyepe Ave St.
- Ajose St St.
- Akanwi St St.
- Akeju St St.
- Alade St St.
- Alafia St St.
- Alhaji Kalejaiye St St.
- Animashaun St St.
- Anuoluwapo St St.
- Apata St. St.
- Araromi Lane St.
- Arujo Crescent St.
- Asubiuro Drive St.
- Awe Crescent St.
- Awoseyin St St.
- Ayoade St St.
- Babalola St.
- Bailey Lane St.
- Bailey St St.
- Bajulaiye Compund St.
- Bajulaiye Rd St.
- Balogun St St.
- Bashua St St.
- Boyle St St.
- Church St St.
- Craig St St.
- Dabira St St.
- Dally St St.
- Durosinmi St St.
- Duru St St.
- Efonlaye St St.
- Emmanuel Kolawole St
- Esan Ogbogun St.
- Eyiowulani St St.
- Fadipe St St.
- Fakorede St St.
- Falola St St.
- Famgero St St.
- Fatai Ade St St.
- Fatai Kadiri St St.
- Fatomi Crescent St.
- Femi Adebule St St.
- Folagoro St St.
- Folami St St.
- Gbadamosi St St.
- George Alade Lane St.
- George St St.
- Gladstone Lane St.
- Hassan Bakare St St.
- Humuani St St.
- Ibrahim Alli St St.
- Ikija St.
- Ikorodu St St.
- Israel Oyekan St St.
- Iyanda Bashua St.
- Jabita Close St.
- Jemilugba St St.
- Johson St St.
- Market St St.
- Martins St St.
- Masalashi St St.
- Modupe St St.
- Morocco Rd St.
- Nnemeka St St.
- Obiwunmi St St.
- Odubiyi St St.
- Odumuyiwa St St.
- Odunlade St St.
- Odunlami St St.
- Ogunbadejo St St.
- Oguntolu St St.
- Okesuna St St.
- Okujobi St St.
- Okuyiga St St.
- Olabiran St St.
- Olaleye St St.
- Olasode St St.
- Olatunde St St
- Olatunji Sanusi St St.
- Olufeko St St.
- Olufemi St St.
- Oluwadere St St.
- Oluwalogbon St St.
- Onafawakan St St.
- Opanubi St St.
- Opeloyeru S St.
- Opere St St.
- Orile St St.
- Owolabi Balogun St St.
- Oyedele St St.
- Party St St.
- Petiti Daho St St.
- Rufai St St.
- Saka St St.
- Sanusi St St.
- Shipeolu St St.
- Shodeke St St.
- Shofiya St St.
- Sodimu St St.
- Teslimiage Close St.
- Tola Oduntan St St.
- Tunji Oluwole Close St.
- Unity Crescent St.
- Vicent St St.
- Watch Towers St.

## REFERENCES

Bates A. (2008) The Development of a 'Postcode Best Fit' Methodology for Producing Population Estimates for Different Geographies *Population Trends* 133 National Statistics

Bartley M. and Owen C. (1996) Relation between Socioeconomic Status, Employment, and Health During Economic Change, 1973 – 93 *British Medical Journal* 313 pp 445 - 449

Boone C. G. (2008) Improving the Resolution of Census Data in Metropolitan Areas Using a Dasymetric Approach: Applications for the Baltimore Ecosystem Study *Cities and the Environment* 1 (1) pp 1 - 25

Birkin M. (1995) *Customer Targeting, Geodemographics and Lifestyle Approaches* Longley P. & Clarke G. (eds.) GIS for Business and Service Planning. Cambridge. Geoinformation International, pp. 104-149.

Coulombe H. and Wodon Q. (2007) Combining Household Survey Data for better Targeting: The West and Central Africa Poverty Mapping Initiative *Poverty Data, Measurement and Policy Special Expanded Edition* 280

DEPARTMENT OF THE ENVIRONMENT (1987) *Handling Geographic Information: Report of the Committee of Enquiry Chaired by Lord Chorley* HMSO, 23.

DETR (2000) Measuring Multiple Deprivation at the Small Area Level: The Indices of Deprivation 2000 *Department of the Environment, Transport and the Regions*

Evans M. (1998) From 1086 and 1984: Direct Marketing into the Millennium *Marketing Intelligence and Planning* 16 (1) pp 56 – 67

Evans N. and Webber R. (1994) Advances in geodemographic classification techniques for target marketing *Journal of Targeting Measurement and Analysis for Marketing* 2 pp 313 --321.

Exeter D. J., Boyle P., Feng Z., Flowerdew R. and Schierloh N. (2005) The Creation of 'Consistent Areas Through Time' (CATTs) in Scotland, 1981 – 2001 *Population Trends* 119 National Statistics

Farr M. and Webber R. (2001) MOSAIC: From an Area Classification System to Individual Classification *Journal of Targeting, Measurement and Analysis for Marketing* 10 (1) pp 55 – 65

Fieldhouse E. and Tye R. (1996) Deprived People or Deprived Places? Exploring the Ecological Fallacy in Studies of Deprivation with the Samples of Anonymised Records *Environment and Planning A* 28 pp 237 - 259

Harris R. and Longley P. (2002) Creating Small Area Measures of Urban Deprivation *Environment and Planning A* 34 pp 1073 – 1093

Harris R. and Longley P. (2004) *Targeting Clusters of Deprivation within Cities* Stillwell J. and Clarke G (eds.) Applied GIS and Spatial Analysis. John Wiley, London pp 89 - 110

Jiang B. (2007) A Topological Pattern of Urban Street Networks: Universality and Peculiarity *Physica A: Statistical Mechanics and its Applications* 384 pp 647 – 655

Jiang B. and Claramunt C. (2004) Topological Analysis of Urban Street Networks *Environment and Planning B: Planning and Design* 31 pp 151 - 162

Kuipers B. (1978) Modelling Spatial Knowledge *Cognitive Science* 2 pp 129 - 153

Longley P. and Harris R. (1999) Towards a New Digital Data infrastructure for Urban Analysis and Modeling *Environment and Planning B Planning and Design* 26 pp 855 – 878

Martin D. (1991) *Geographic Information Systems and their Socioeconomic Applications* Routledge, London

Martin D. (1992) Postcodes and the 1991 Census of Population: Issues, Problems and Prospects *Transactions of the Institute of British Geographers, New Series* 17 (3) pp 350 – 357

Mennis J. and Hultgren T. (2006) Intelligent Dasymetric Mapping and its Application to Areal Interpolation *Cartography and Geographic Information Science* 33 (3) pp 179 - 194

NIPOST (2003) *Nigerian Postcode Directory, 2ND Edition* FOMAT Press, Lagos

Oladeji S. and Abiola A. (2000) Poverty Alleviation with Economic Growth Strategy: Prospects and Challenges in Contemporary Nigeria *Journal of Social Development in Africa* 15(2) pp 33 – 53

O'Malley L., Patterson M. and Evans M. (1995) Retailing Applications of Geodemographics: A Preliminary Investigation *Marketing Intelligence and Planning* 13 (2) pp 29 - 35

O'Malley L., Patterson M. and Evans M. (1997) Retailer Use of Geodemographics and other Data Sources: An Empirical Investigation *Journal of Retail and Distribution Management* 25 (6) pp 188 - 196

ONS (2005) Beginners Guide to UK Geography: Postal Geography
*Available at: http://www.statistics.gov.uk/geography/postal_geog.asp.* Accessed: 17/09/2009

Prothero R. (1963) African Population Maps: Problems and Progress *Geografiska Annaler* 45(4) pp 272 – 277

Porta S., Crucitti P. and Latora V. (2006) The Network of Urban Streets: A Dual Approach *Physica A: Statistical Mechanics and its Applications* 369 pp 853 - 866

Raper J. F., Rhind D. W. and Shepherd J. W. (1992) *Postcodes: The New Geography* Longman Scientific and Technical, London.

Vickers D. and Rees P. (2006) Introducing the Area Classification of Output Areas *Population Trends* 125 National Statistics

Vickers D. and Rees P. (2007) Creating the UK National Statistics 2001 Output Area Classification *Journals of the Royal Statistical Society A* 170 (2) pp 379 – 403

Webber R. (1985) The Use of Census Derived Classifications in the Marketing of Consumer Products in the United Kingdom *Journal of Economic Measurement* 13 pp 113 – 124

Webber R. (2004) Designing geodemographic classifications to meet contemporary business needs *Interactive Marketing* 5 (3) pp. 219-237.

Winkleby M., Jatulis D., Frank E. and Fortmann S. (1992) Socioeconomic Status and Health: How Education, Income , and Occupation Contribute to Risk Factors for Cardiovascular Disease *American Journal of Public Health* 82 (6) pp 816 - 820

Wong D. (2009) *The Modifiable Areal Unit Problem (MAUP)* Fotheringham A. S. and Rogerson P. A. (eds). The SAGE Handbook of Spatial Analysis. SAGE Publications, London pp 105 - 124

**BIOGRAPHY**

*Nicholas, a third year PhD researcher at Kingston University's Centre for GIS, is interested in geodemographics, geomarketing and its application in developing countries; he believes the insight it would offer on a developing country's population will provide a better framework for marketing and infrastructure decisions within such areas.*

# How Many Volunteers Does It Take To Map An Area Well?

Mordechai (Muki) Haklay, Aamer Ather, Sofia Basiouka

Department of Civil, Environmental and Geomatic Engineering,
UCL, Gower St., London, WC1E 6BT
Tel. +44 20 7679 2745 | Email: m.haklay@ucl.ac.uk

KEYWORDS: Volunteered Geographical Information, OpenStreetMap, Linus' Law

## 1. Introduction

Of the range of geographical information technologies that emerge in the era of Web Mapping 2.0 (Haklay, Singleton and Parker, 2008), the applications that involve geographical information collection and sharing by a wide range of participants deserve special attention, as they present a significant departure from past practices. Goodchild (2007) coined the term 'Volunteered Geographic Information' or VGI for this phenomena. This is rather unfortunate term, as not all the data is volunteered knowingly and explicitly. A better generic term for the phenomena could focus on the mode of production and following Howe's (2006) definition of crowdsourcing call this Crowdsourced Geographical Information, or describe it factually as user-generated spatial content (Antoniou, Haklay and Morley, in press).

However, there is one class of systems that sparked specific interest as it is indeed based on volunteered information in the common sense of the word. These systems allow people from across the globe to create and share geographical content in an egalitarian manner, in what Benkler and Nissenbaum (2006) entitle 'Commons-Based Peer Production'. OpenStreetMap (OSM), which started at UCL in 2004, is a chief example of such a system. In OSM, volunteers (currently over 180,000) contribute to create a free editable vector map of the world (see Haklay and Weber, 2008, for a detailed discussion). While many of the early volunteers of OSM were highly technically literate, they were not necessarily experts in geographical data collection. As the project increases in size and provides new tools such as 'Walking Papers' (Migurski, 2009), which allow participants to print paper maps and use them for data collection activities, no assumption can be made about the background knowledge or data collection ability of volunteers.

In light of the data collection by amateurs, the distributed nature of the data collection and the loose coordination among them, 'how good is the quality of the information that is collected through such activities?' becomes a significant question. This is a crucial question about the efficacy of VGI activities and the value of the outputs for a range of applications, from basic navigation to more sophisticated applications such as site location planning.

Research by Haklay (in press), Ather (2009) and Kounadi (2009) has demonstrated that, in terms of positional accuracy, the quality of OSM data is comparable to traditional geographical datasets that are provided by national mapping agencies (Ordnance Survey in the UK and the Hellenic Military Geographical Service in Greece). These studies were based on the process that Goodchild and Hunter (1997) developed, which provides an estimation of the overlap between a reference dataset that is assumed to have a higher positional accuracy and a test dataset. In all these comparisons, the national mapping agency was used as the reference dataset and OSM as the test dataset. The results show overlap of about 80% in most cases.

However, because the information is provided by a range of participants, there is a need to understand at what stage of the data collection process the quality of the data becomes reliable. This is the topic of this paper.

## 2. Evaluating Linus' Law for OpenStreetMap

OSM is not the first commons-based peer-production activity. There are, in fact, many projects similar to it, albeit not in the area of geographical information. Many (though not all) open source projects have similar characteristics in terms of the distributed development effort and loose coordination amongst participants. Thus, parallels can be drawn between the quality issues of VGI and questions about code and software quality that were raised about many open source projects by the mainstream commercial software development community such as the development of GNU software, or Linux operating system (Raymond, 2001).

While there are many mechanisms through which open source projects ensure the quality of the software (Halloran and Scherlis, 2002), the number of people that are engaged in it receive special attention in what is known as Linus' Law. The law states that 'Given enough eyeballs, all bugs are shallow' (Raymond, 2001, p.19) and means that because in open source project the underlying code is open for scrutiny by any person, when a bug occurs in the code, it will be found and fixed because many people look at the code. For mapping, this can be translated into the number of contributors that worked on a given area. The rationale behind it is that, if there is only one contributor in an area, he or she might inadvertently introduce some errors. For example, they might forget to survey a street or might position a feature in the wrong location. However, if there are several contributors, they might notice inaccuracies or 'bugs' and, therefore, the more users, the fewer 'bugs'.

If Linus' Law applies to VGI, it could provide an easy-to-calculate method for quality evaluation – for example, as in Haklay (in press), a regular grid can be used to count the number of contributors per square kilometre as a proxy for accuracy and can assist in decisions about fitness for use.

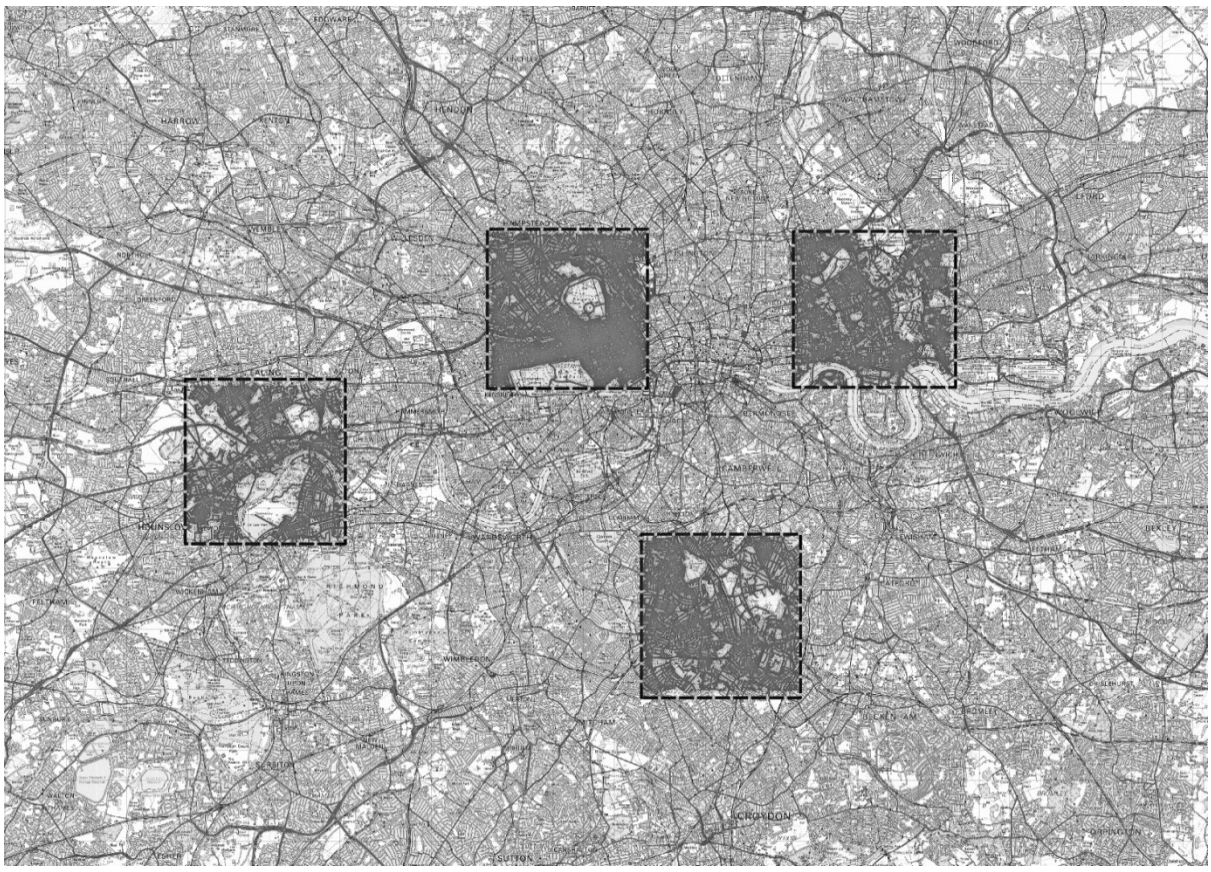


**Figure 1:** ITN data used for the comparison across London

The evaluation of Linus' Law's relevance to OSM was carried out by comparing the positional

accuracy of OSM in 125 square kilometres of London (Figure 1). A detailed comparison of OSM and the Ordnance Survey MasterMap Integrated Transport Network (ITN) layer was carried out first (Ather 2009). The results of this study were divided for each grid square, so it was possible to calculate the overall positional accuracy estimation for each cell. The value is a weighted average of the overlap between OSM and OS objects, weighted by the length of the ITN object. The next step was to compare the results to the number of users at each grid square, as calculated from the details of user name, taken form information that is linked to nodes in the area.

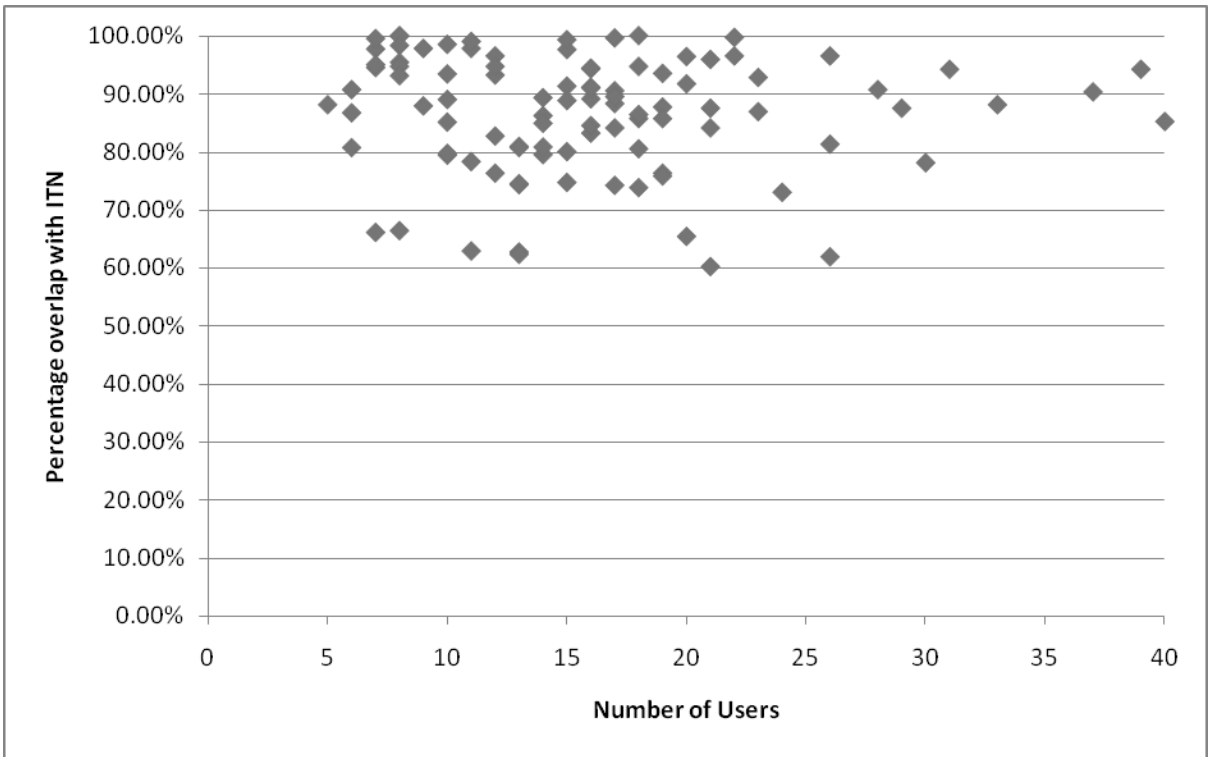



**Figure 2:** Number of OSM contributors and estimated positional accuracy

The results (Figure 2) show that, above 5 users, there is no clear pattern of improved quality. The graph shows that the quality, while generally very high, is not dependent on the number of users – so Linus' Law doesn't apply to OSM (and probably not to VGI in general) above 5 users.

From other studies of OSM data, the explaining hypothesis is that, due to the participation inequality in OSM contribution (some users contribute significant amount of information while others contribute very little), the quality is actually linked to a specific user, and not to the number of users. However, this hypothesis need to be tested.

Yet, further research is necessary. Firstly, the analysis was carried out in London – so it is necessary to evaluate data quality and number of contributors in other parts of the country where different contributors have collected the data, a preliminary analysis of an automatic method to evaluate positional accuracy across England shows similar trends. Secondly, the analysis did not include the interesting range of 1 to 5 users, and it might be the case that there is rapid improvement in quality from 1 to 5 and then it doesn't matter. Maybe the big change is from 1 to 3? Finally, the analysis focused on positional accuracy, and it is worth exploring the impact of the number of users on completeness.

## 3. Acknowledgements

The 1:10,000 raster and the SOA boundaries were provided under EDINA/Digimap and EDINA/UKBorders agreements.

## References

Antoniou, B., Haklay, M., and Morley, J. (in review) Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon, submitted *to Geomatica special issue on VGI*.

Ather, A. (2009) 'A Quality Analysis of OpenStreetMap Data' unpublished M.Eng. dissertation, UCL

Benkler, Y., and Nissenbaum, H. (2006) Commons-Based Peer Production and Virtue. *Journal of Political Philosophy*, 14 (4): 394-419

Goodchild, M. F., and Hunter, G. J. (1997) A simple positional accuracy measure for linear features, *International Journal of Geographical Information Science*, 11(3): 299-306

Goodchild, M.F. (2007) Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0, *International Journal of Spatial Data Infrastructures Research*, 2: 24-32

Haklay, M. (in press) How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England, *Environment and Planning B*

Haklay, M., and Weber, P. (2008) OpenStreetMap – User Generated Street Map, *IEEE Pervasive Computing*. October-December 2008, 12-18

Haklay, M., Singleton, A., and Parker, C. (2008) Web Mapping 2.0: the Neogeography of the Geoweb, *Geography Compass* 3: 2011-2039, doi: 10.1111/j.1749-8198.2008.00167.x

Halloran T.J., and Scherlis W.L. (2002) *High quality and open source software practices. Meeting Challenges and Surviving Success.* 2nd Workshop on Open Source Software Engineering, University College Cork, Ireland, 2002. Available at: http://opensource.ucc.ie/icse2002

Howe, J. (2006) The Rise of Crowdsourcing, *Wired Magazine*, June 2006

Kounadi, O. (2009) 'Assessing the quality of OpenStreetMap data', unpublished MSc dissertation, UCL

Migurski, M. (2009) *Walking Papers* (walking-papers.org)

Raymond, E.S. (2001) *The Cathedral and the Baazar*, O'Reilly

## Biography

*Mordechai (Muki) Haklay is a senior lecturer in Geographical Information Science in the department of Civil, Environmental and Geomatic Engineering at UCL. His research interests are in public access to environmental information, Human-Computer Interaction (HCI) and Usability Engineering for GIS, and societal aspects of GIS use.*

*Aamer Ather holds an MEng in Geoinformatics from University College London. For his final year thesis he investigated the quality and accuracy of OpenStreetMap data using GIS software. After Graduating in 2009, he is now exploring various career opportunities and is currently in employment with Infotech, working on a project with the Rural Payments Agency.*

*Sofia Basiouka is a PhD student in the department of Rural and Survey Engineering at the National Technical University of Athens (N.T.U.A). She recently completed the MSc in GIS at UCL (2009) and the MEng in Survey Engineering at N.T.U.A.(2008). Her research is focused on Cadastre.*

# A step towards the improvement of spatial data quality of Web 2.0 geo-applications: the case of OpenStreetMap

Vyron Antoniou[1], Mordechai (Muki) Haklay[2]
Jeremy Morley[3]

[1,2]University College London, Gower Street, London, WC1E 6BT, UK
+44 20 7679 2745
{v.antoniou, m.haklay}@ucl.ac.uk
[3]University of Nottingham, University Park, Nottingham, NG7 2RD, UK
+44 (0) 115 84 68411
Jeremy.Morley@nottingham.ac.uk

KEYWORDS: VGI, OSM, Spatial Data Quality, Quality Assurance, Vector Data

## 1. Introduction

OpenStreeMap (OSM) is a Web 2.0 system that allow users to create and freely use spatial data. The success of OSM initiative has drawn the attention of scholars and researchers that started to examine issues like the credibility, quality and fitness for purpose of such data (Goodchild 2007, Sui 2008, Flanagin and Metzer 2008, Haklay 2008). In this study we also look into data quality issues (mostly focusing on the attribution of the entities), but from a different perspective. We analyze and formalise the knowledge submitted to the wiki pages by the contributors of OSM regarding the map creation process. This set of rules functions as a user guide for the creation of spatial data for OSM and it can be considered as the 'specification' of the OSM geodata product. After explaining the nature of these specifications, they are used to evaluate the quality of the data created.

## 2. The OSM Rules

There is a widespread manifestation in the OSM wiki pages that OSM community does not want to impose any rules on its participants. On the contrary, the wiki pages claim that participants can freely use any lawful method and practice to create spatial content and also that they are free to assign any kind and type of attributes (using tags) to real world features (OSM 2009):

*"OpenStreetMap does not have any content restrictions on tags that can be assigned to Nodes, Ways or Areas. You can use any tags you like".*

In practice though, OSM users have created numerous wiki pages that are full of instructions regarding procedures to describe geographical objects (OSM 2009):

*"However, there is benefit in agreeing on a recommended set of features and corresponding tags in order to create, interpret and display a common basemap".*

These instructions are not presented as hard and fast rules but rather as lessons from other contributors' experiences or as best practice proposals. Nonetheless, this wiki-made user guide has evolved into a quite complicated and some times hard to follow technical document.

It is interesting to note that the road map to create or change such a rule is totally democratic. In brief, users can start a proposal procedure whenever they feel that a mapping feature should be added or changed. This procedure includes a discussion and a voting step which determines whether the proposal will be rejected or accepted and consequently implemented. The active and approved map features are documented with proper instructions and both written and visual examples. This is a continuous process; entities from the map features list can be replaced with new ones and the old entities become deprecated.

## 2. Quality assurance and OSM

This open and democratic process perhaps is one of the key factors for the popularity of the OSM endeavour. However, such freedom created a lot of inconsistencies and therefore there was a need for some form of quality assurance mechanism that would enable users to correct inaccuracies. Indeed, today there is a variety of options for an OSM contributor to achieve that: from assigning a simple "fixme" tag to a feature to indicate that it requires updating to using one of the dedicated applications for identifying errors in the database (the list of those applications is available at http://wiki.openstreetmap.org/wiki/Quality_Assurance).

These early attempts for identifying and correcting mistakes in the OSM dataset, while they present interesting paradigms of a self-correcting mechanism for a crowd–sourced, Web 2.0 application, are still incomplete and patchy. The 'Keep Right' application that was created by the OSM community and monitors the violation of some OSM rules provides an example for this (Figure 1). The application evaluates the data conformance against a set of pre-defined rules and presents to the users the positions of possible mistakes.

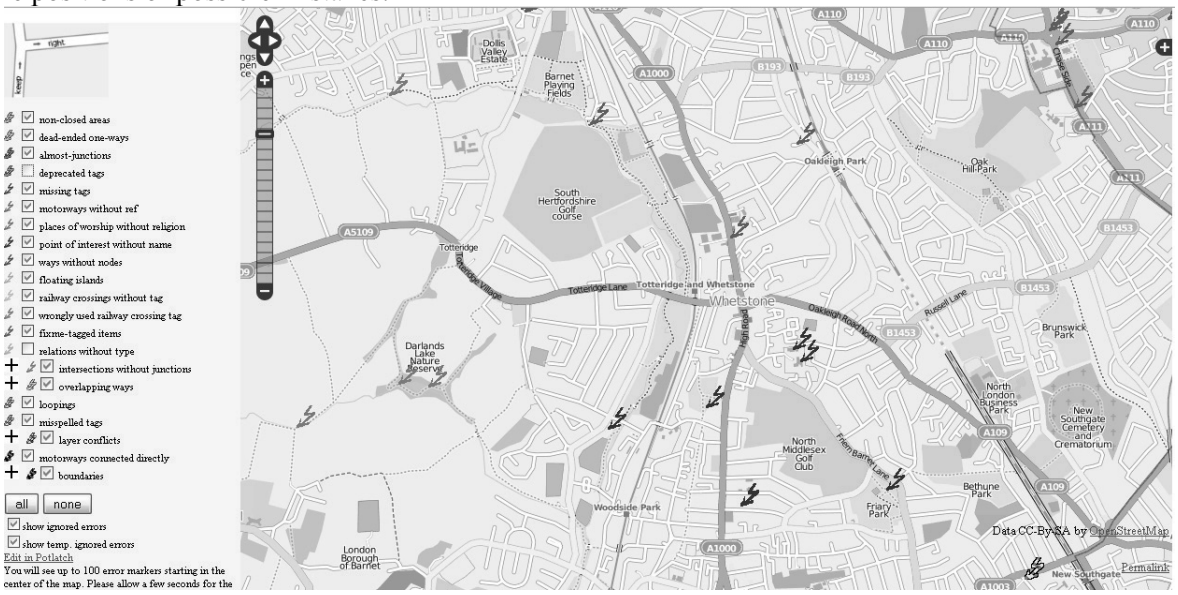



**Figure 1.** The Keep Right application. The table of contents in left hand side contains the rules monitored by the application.
(Source: Keep right 2010)

The question that needs to be answered at this point is *"How these specific rules have been chosen?"* and most importantly *"Why only these rules? Do they provide a quality assurance mechanism for the entire OSM dataset? If no, what should we do? ".* In fact, these early efforts add to the argument that there is a strong need for a more holistic approach when it comes to building quality assurance mechanisms and determining the spatial data quality for data generated in GeoWeb applications.

Nonetheless, from the standing point of someone who wishes to use the OSM data there is no clear indication regarding the quality of that product. This stems from the fact that there are no solid specifications that will help to determine the magnitude of the product conformance. Therefore, we propose an XML Schema able to model the OSM entities as described in the Map Features list (http://wiki.openstreetmap.org/wiki/Map_Features). For example, Figure 2a shows the object type of an abstract OSM object according to the existing rules, while Figure 2b shows the object type of an abstract Motorway object.

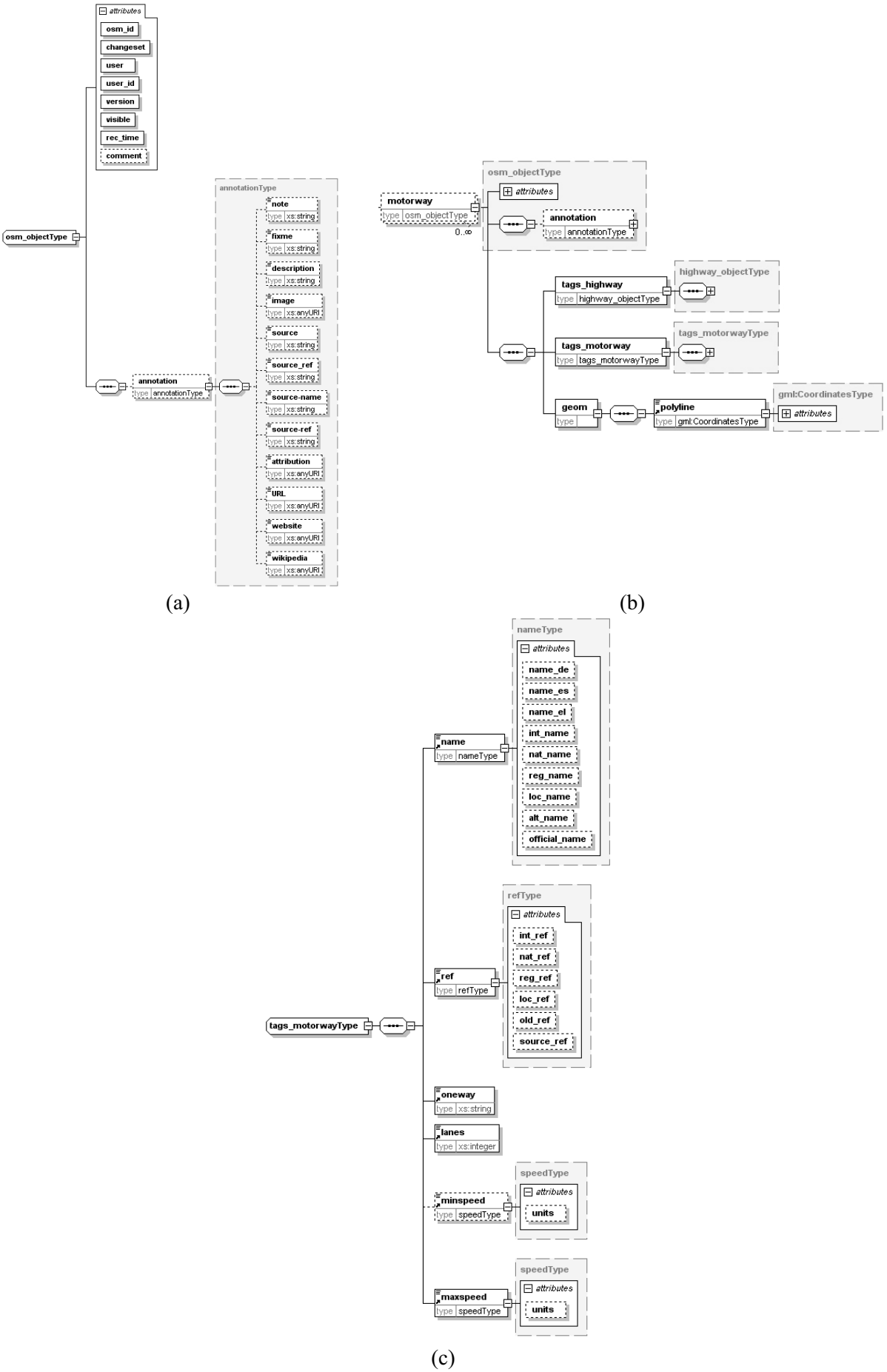


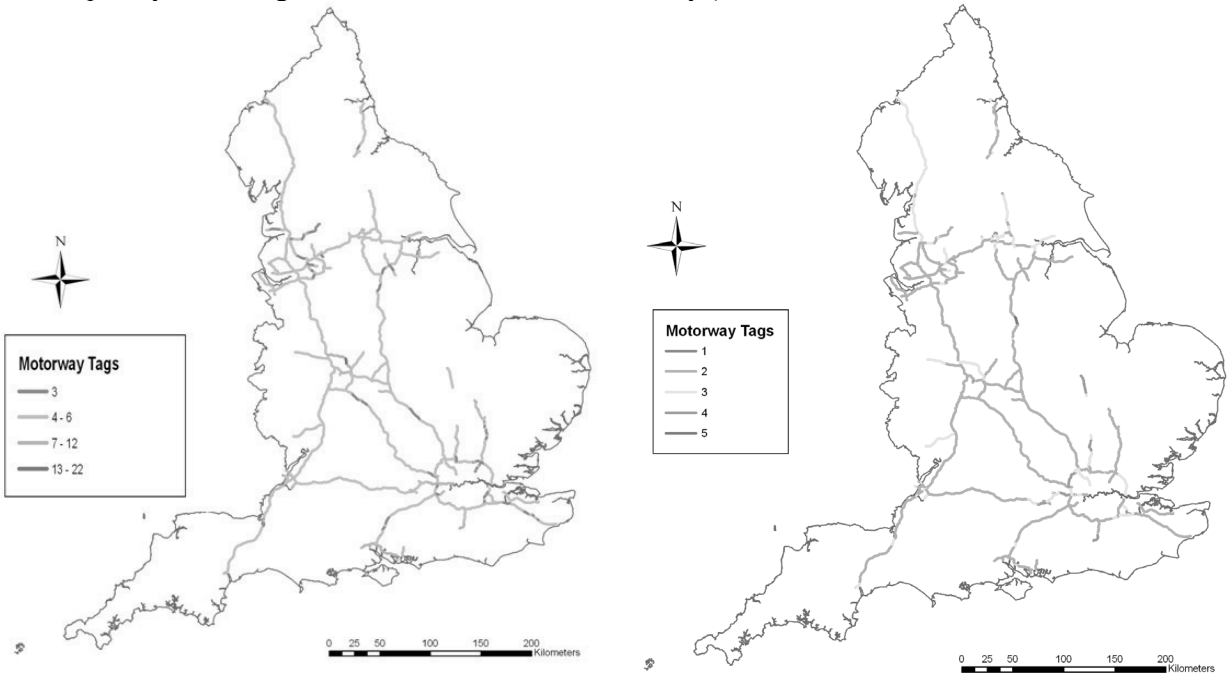

**Figure 2.** (a) The XML Schema fragment that shows the attributes of an abstract OSM object (b) The XML Schema fragment for motorways, (c) The XML Schema of the motorway tags.

It is important to emphasize that this XML schema has been constructed by following the rules that are described in the OSM wiki pages. Consequently, this XML Schema cannot be regarded as a static document but rather as a set of rules that is constantly synchronized with the rule creating system of OSM.

By formalizing and applying a common language to the rules that affect the creation of the OSM map features it is possible to analyze and measure the level of conformity or violation of the OSM database against those rules. We used the XML schema (and more specifically the XML fragment of each entity as shown for example in Figure 2b for Motorways) in order to examine the quality of OSM features for England. In this effort we followed the methodologies documented in the ISO 19113 (ISO 2002) and ISO 19114 (ISO 2003) standards.

## 4. Findings and conclusions

Our findings show that wherever there is a systemic way (either a database schema or elements in the user interface of OSM editing applications such as JOSM and Potlach) to handle the data input or to guide user contribution, then the data quality is high. On the contrary, in the cases where there is no guidance other than the wiki pages the quality is considerably lower. For example, Figure 3a shows the total number of tags assigned to each motorway while Figure 3b shows the number of XML Schema conforming tags that have been assigned to motorways. It can be seen that while OSM contributors have assigned up to 22 tags to a single motorway element, in fact for the majority of the motorways only 2 to 3 tags are valid according to the guidelines published at the wiki pages (and consequently according to the XML Schema for motorways).




(a) (b)

**Figure 3.** (a) The number of XML Schema conforming tags for each of the motorway entities in England and (b) the total number of tags assigned to each entity.

In other words, it is not uncommon to witness the violation of OSM rules, despite the fact that all these rules have been established with the open and democratic procedure which was described earlier. Finally, it should be noted that since data quality changes whenever there is a change in the data (e.g. due to a transformation), in the ground truth or in the specifications of the product, we have witnessed that constant changes of OSM specifications through the voting system affects negatively the overall quality of the dataset. For example, before the decision to deprecate the “created_by” tag

from all the entities in OSM the 22.44% of the Motorways in England violated the conceptual consistency of the XML Schema. This figure climbed to 80.62% when the rule and thus the XML Schema changed.

Therefore, we conclude that for GeoWeb applications, the first step towards the improvement of their data is firstly to conceptually formalize the data sought. This step will enable them to create the necessary environment both in the front- and in the back-end of the application so to help contributors to conform to the chosen schema. In turn, this will improve data quality by diminishing errors and inconsistencies in the dataset. Additionally, whenever there is a change in the specifications the formalisation can facilitate automatic correction of the existing data.

Finally, it is important to note that, such a process does not affect in any way the openness of a Web 2.0, crowd-sourcing application or the excitement and sense of freedom that volunteer geographers feel when contributing to such applications. By doing so, we can have both the openness and the formalization needed to achieve a crowd-sourced dataset of high quality.

## References

Flanagin A., Metzer M., 2008. The credibility of volunteered geographic information. GeoJournal,72: 72:137–148.

Goodchild, M. F., 2007. Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), 211–221.

Haklay, M., 2008. How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. [Online] Available at: http://www.ucl.ac.uk/~ucfamha/OSM data analysis 070808_web.pdf [Accessed 20 November 2009]

ISO (International Standardisation Organisation), 2002. ISO 19113:2002 Geographic information — Quality principles.

ISO (International Standardisation Organisation), 2003. ISO 19114:2003 Geographic information — Quality evaluation procedures.

Keep right. 2010. Data consistency checks for OSM [Online] Available at: http://keepright.ipax.at/ [Accessed 20 February 2010].

OSM. 2009. Map Features. [Online] (Updated 12 Oct 2009) Available at: http://wiki.openstreetmap.org/wiki/Map_Features [Accessed 20 November 2009].

Sui, Z., D., 2008. The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. In Computers, Environment and Urban Systems 32 :1–5.

## Biography

*Vyron Antoniou is a Captain in the Greek Army and since 1998 serves at Hellenic Military Geographical Service. Currently is a PhD student at UCL in the department of Civil Environmental and Geomatic Engineering. His research interests are in user generated content and National Mapping Agencies; spatial databases; vector data transmission over the Web and web mapping.*

*Mordechai (Muki) Haklay is a senior lecturer in Geographical Information Science in the department of Civil, Environmental and Geomatic Engineering at UCL, where he is also the Director of the Chorley Institute, which focuses on facilitating interdisciplinary geospatial research at UCL. His research interests are in public access to environmental information, Human-Computer Interaction (HCI) and Usability Engineering for GIS, and Societal aspects of GIS use. He received his PhD in Geography from UCL.*

*Jeremy Morley is Deputy Director of the Centre for Geospatial Science at the University of Nottingham and was previously a lecturer in Geographic Information Systems at University College London. His interests lie in GIS interoperability; interfaces between GIS web services and mashup technology; sensor webs; and planetary mapping."*

# Automatically generating keywords for georeferenced images

Ross S. Purves[1], Alistair Edwardes[1], Xin Fan[2], Mark Hall[3] and Martin Tomko[1]

[1] Department of Geography, University of Zurich, Switzerland
ross.purves@geo.uzh.ch; ali.edwardes@gmail.com; martin.tomko@geo.uzh.ch
[2] Department of Information Studies, University of Sheffield, UK
x.fan@sheffield.ac.uk
[3]Department of Computing Science, Cardiff University, UK
M.M.Hall@cs.cardiff.ac.uk

KEYWORDS: georeference, image, keyword, indexing, caption

## 1. Introduction

The possibility of automatically annotating information objects such as images with metadata related to their geography is increasingly relevant, as larger volumes of such information are captured with explicit georeferences (e.g. Flickr images). Within the Tripod project research has been carried out as to the extent to which we can automatically derive keywords and captions for images which are captured by cameras capable of automatically storing coordinates and azimuth information, that is to say not only the position *from* which a picture was taken, but also the direction in which the camera was pointing. Improved tagging of images can improve search, and particularly in commercial contexts, reduce annotation costs which are an important part of the work of image libraries (Iwasaki et al., 2008).

In previous work we have described how the Pansofsky-Shatford facet matrix (Table 1) formalises ways in which we might consider carrying out this task (Purves et al., 2008). The matrix has three levels, termed the *specific of*, the *generic of* and *about*. Each level has four associated facets: *who*, *what*, *where* and *when*. In GIScience the *where* facet is of particular interest, with the *where/ specific of* element referring to how we name locations, in other words the toponyms that we assign to a particular location. The *where/ about* element relates to the qualities that a location might manifest for us, thus, for example, the degree of remoteness or the warmth conveyed by an image. The subject of this paper is though the *where/ generic of* element, that is to say generic concepts such as *church*, *village* and *beach*.

**Table 1.** The Pansofsky-Shatford facet matrix (Shatford (1986), p. 49)

| ***Facets*** | *Specific Of* | *Generic Of* | *About* |
|---|---|---|---|
| *Who?* | Individually named persons, animals, things | Kinds of persons, animals, things | Mythical beings, abstraction manifested or symbolised by objects or beings |
| *What?* | Individually named events | Actions, conditions | Emotions, abstractions manifested by actions |
| *Where?* | Individually named geographic locations | Kind of place geographic or architectural | Places symbolised, abstractions manifest by locale |
| *When?* | Linear time; dates or periods | Cyclical time; seasons, time of day | Emotions or abstraction symbolised by or manifest by |

In this paper we set out to describe the stages involved in automatically generating keywords for georeferenced images, and illustrate the methods applied through a small number of examples, before discussing further work which we will report at GISRUK.

## 2. Methods

In order to assign keywords to images based on location, a number of calculations must be carried out. Within the Tripod project we first identify the viewshed of an image, based on the camera parameters, camera location, azimuth and terrain in rural areas. In urban areas the camera location is simply buffered using an empirically derived distance to form an image sector. Having generated a visible geometry, we then sample a variety of spatial datasets and explore the classes found within these datasets. Since different datasets have very different contents, we use a *concept ontology* (Edwardes et al., 2007) to map between dataset instances and concepts that are used by keywords. Lastly, the complete list of candidate concepts is ranked and filtered to give a final set of candidate keywords.

### 2.1 Identifying visible area

The first stage in identifying the visible area for a particular camera is the extraction of device metadata from the EXIF header of the image. This metadata contains information about the camera settings (e.g. focal length) and is used to determine, for example, the angular width of the visible area. We first identify urban images using landcover data, and then assume that in such regions if a building is visible in the image, it is proximal and assign a buffer of 50m to the sector identified.
We identify buildings in images using a simple content-based building detection algorithm.

In rural areas, where terrain plays a much more important role in specifying the visible area, we calculate the viewshed for the angularly restricted sector identified from the camera metadata using standard methods and Shuttle Radar Topography Mission (SRTM) DEM data with a resolution of 90m (Figure 1). In previous work we explored the sensitivity of the buffer size and the distance over which viewsheds were calculated, particularly with respect to visibility of point-like objects and found that thresholds of around 1000m were sensible for objects with a width of ~25m (Tomko et al., 2009). Our approach assumes that urban viewsheds are limited by objects, rather than terrain, which is clearly an oversimplification in some cases.

### 2.2 Linking spatial data to concepts

The underlying hypothesis in Tripod is that spatial data will describe many aspects of an image taken at some location, assuming that its focus is somehow geographic. A key challenge is linking multiple datasets to concepts in an extensible way which allows the integration of multiple datasources, and generates keywords which are not specific to individual datasets. The use of different datasets allows the assignment of different types of concepts at different scales. Thus, for example, we can integrate land cover data and topographic data from different National Mapping Agencies or OpenStreetMap.

Keywords are assigned by mapping individual items in a dataset to concepts in an ontology, which was generated by exploring user-generated content such as Geograph and contains a range of relationships between concepts (Edwardes et al., 2007). Thus, multiple dataset items might map to the same concept and we can introduce new data at any time to the system.

We identify candidate concepts by, in the case of topographic data, producing very high resolution raster representations (typically ~1-5m) where individual footprints are based on estimates of the real-world size of individual classes of objects. All dataset items are assigned unique values, and by intersecting visible areas with concept representations, a list of candidate concepts and their relative areas with respect to the visible area are generated (Figure 1).


**Figure 1.** Spatial data for a visible area corresponding to the second image shown in Figure 2 – ©SwissTopo 1:200 000 and Corine data are shown here – note the viewshed is restricted to a range of 1.5km and is calculated using SRTM 90m data

### 2.3 Ranking and filtering candidate concepts

The simplest way to rank concepts would be simply to use their relative areas. However, this ignores several important aspects. Firstly, not all data are spatially contiguous, and thus landcover-derived concepts will automatically float to the top of any such ranking. Secondly, the *salience* of a particular concept in a scene is likely to be related to not only its spatial footprint, but also its overall rarity with respect to the scene, and thus its *descriptive prominence* (Tomko and Purves, 2009). Thus, a tree in the Sahara desert should be assigned more weight than a tree in central Switzerland. Finally, the web provides us with a potential means of assessing how commonly a particular concept is used in a region. By querying the web with toponyms assigned to the visible area through the process of reverse geocoding (e.g. Smart et al., 2009) and concepts, we can explore the *web prominence* of individual keywords. In the final ranking of concepts, we rank according to area, web prominence and descriptive prominence, filtering ubiquitous concepts which are not commonly used. By combining web and descriptive prominence, we reduce the importance of rare but uninteresting or unphotographed concepts.

## 3. Exemplar results and discussion

Figure 2 illustrates results from the Tripod system for two exemplar images. Keywords were generated in the urban Edinburgh case using Corine landcover data and OpenStreetMap. In the Swiss rural case, keywords were generated using Corine landcover data and a SwissTopo 1:200000 dataset.

college, byway, village

loch, meadows, moor, agriculture, shore, reservoir, forest

**Figure 2.** Two images and top-ranked keywords – the second image corresponds to Figure 1

For the first image, 2 out of the 3 keywords are appropriate (the image shows a school in Edinburgh) whilst the third is extracted from landcover data, where the possible concepts for discontinuous urban fabric are suburbs and village. In this case, village is clearly inappropriate. For the second image (of a lake in Switzerland) the set of keywords agree well with the image, apart from the somewhat incongruous use of loch. This is because loch's German meaning (hole) which causes it to be highly ranked with local toponyms by the web prominence algorithm.

The complete system is designed to automatically generate candidate keywords for images which can be used in both indexing and search. Current work is evaluating the quality of these keywords for large collections, and will be reported on at GISRUK.

**5. Acknowledgements**

This research reported in this paper is part of the project *TRIPOD* supported by the European Commission under contract 045335. We would also like to gratefully acknowledge contributors to Geograph British Isles, see http://www.geograph.org.uk/credits/2007-02-24, whose work is made available under the following Creative Commons Attribution-ShareAlike 2.5 Licence (http://creativecommons.org/licenses/by-sa/2.5/). Many thanks to the referees for their constructive comments which have improved this paper.

**References**

Edwardes, AS, Purves RS, Simone Bircher and, Christian Matyas. (2007). Deliverable 1.4: Concept ontology experimental report. Available at http://tripod.shef.ac.uk/outcomes/public_deliverables/Tripod_D1.4.pdf

Iwasaki, K, Kanbara, M, Yamazawa, K & Yokoya, N (2008). Construction of extended geographical database based on photo shooting history. In *Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval,* ACM, New York, NY, pp. 185-194.

Purves, RS, Edwardes, AJ & Sanderson, M. (2008). Describing the Where – improving image annotation and search through geography. *First Intl. Workshop on Metadata Mining for Image Understanding (MMIU 2008)*.

Smart P, Twaroch F, Tomko M and Jones, C (2009) Deliverable 6.5: Final Toponym Ontology Prototype. Available at http://tripod.shef.ac.uk/outcomes/public_deliverables/Tripod_D6.5.pdf

Shatford S (1986) Analyzing the subject of a picture: a theoretical approach. *Catalogue and Classification Quarterly* pp39–62.

Tomko, M, Trautwein, F & Purves, RS (2009) Identification of Practically Visible Spatial Objects in Natural Environments. In *Proceedings of AGILE 2009*, Springer-Verlag, Vienna, Austria.

Tomko, M, & Purves, RS (2009) Venice, City of Canals: Characterizing Regions through Content Classification. *Transactions in GIS*, pp295-314.

**Biographies**

*Ross Purves is a lecturer in Zurich. Mark Hall is a PhD student in Cardiff. Xin Fan and Martin Tomko are postdoctoral researchers in Sheffield and Zurich respectively, Alistair Edwardes has moved to new pastures in the Department of Communities and Local Government.*

# Micro-blogging mashups: extending the value of social networks through spatial representation

Kenneth Field[1], James O'Brien[1]

[1]Kingston University London, Centre for GIS, Kingston upon Thames, London KT1 2EE
Tel. +44 (0)20 8417 2541
Email Kenneth.field@kingston.ac.uk
http://www.kingston.ac.uk/centreforgis

KEYWORDS: micro-blogging, visualization, mashups, neo-geography

## 1. Introduction

The growth in social networking via blogs, micro-blogs and online forums has given rise to an increasingly interconnected world. There exists a growing desire for maps defining our place, locating us in unfamiliar surroundings, navigating us and providing spatial context for our activities. Instead of remaining spatially anonymous in online environments, people reveal locations for their identity and activities (Gibin *et al.* 2008). This spatial expression illustrates the importance of geography and how maps can add value to a virtual profile, yet the cartography is often lacking which renders the information one-dimensional (Wood *et al.* 2007).

We explore this spatial expression using the social networking tool Twitter which permits micro-blog postings (tweets). Examples of 'Twitter maps' are reviewed which utilize the Twitter API in concert with map services to spatially represent elements of the tweet. These maps are cartographically simplistic and little work has been done to explore how researchers might capitalise on the spatial expression or how mashups might be designed to reveal more than the simple representation of location.

This paper explores the value and purpose of taking a geographical approach to better represent the micro-blogging environment created by Twitter. It offers two Twitter Map case studies to explore how the spatial component can be accessed and used meaningfully in different ways, firstly to support online asynchronous collaborative learning and, secondly, to represent multiple tweets using innovative cartographic representations as a means of making information more visible.

## 2. Background

Developments brought about by Web 2.0 have reinvigorated cartography and provided technologies that support a rich and diverse mapping landscape (Miller, 2006; Crampton, 2009)). The availability of online map and data services and the ability for non-experts to author online spatial content, provides a cartographic challenge for effective information organisation and visualization (Haklay, 2008; Graham, 2009).

Since its creation in 2006, Twitter has become a hugely popular online social networking tool that allows users to post tweets of 140 characters in a similar fashion to SMS messaging. Twitter is

undeniably a hugely successful information network yet the spatial aspect of the data has been largely overlooked in a meaningful sense and the only visible spatial expression has been in an individual's profile (Field, 2009). A number of imaginative visualisations have been offered that go beyond the simple twitter timeline (e.g. Clark 2008; Schmidt 2009), few have harnessed the locational component.

Embedding certain information within a tweet has allowed others to capitalise on locational information in different ways. Using #hashtags allows people to embed key words into a tweet which can be used as an organising framework. These frameworks can be subsequently explored or visualized collectively where numerous people are tweeting about a similar subject. One good example of this was during the snowfall on 1st February 2009 in the United Kingdom. As snow began to fall a #uksnow tag took hold on Twitter providing real time reports of snowfall where the tweet had "#uksnow", a UK postcode (first part only) and optionally a mark in the form x/y indicating how hard it was snowing. Figure 1 illustrates an example by Darbyshire (2009) that makes use of the Twitter and Google APIs to map location (from the Twitter profile) and intensity (based on the 1-10 user-defined rating). Refinements of this work have been produced by Marsh (2009, 2010).

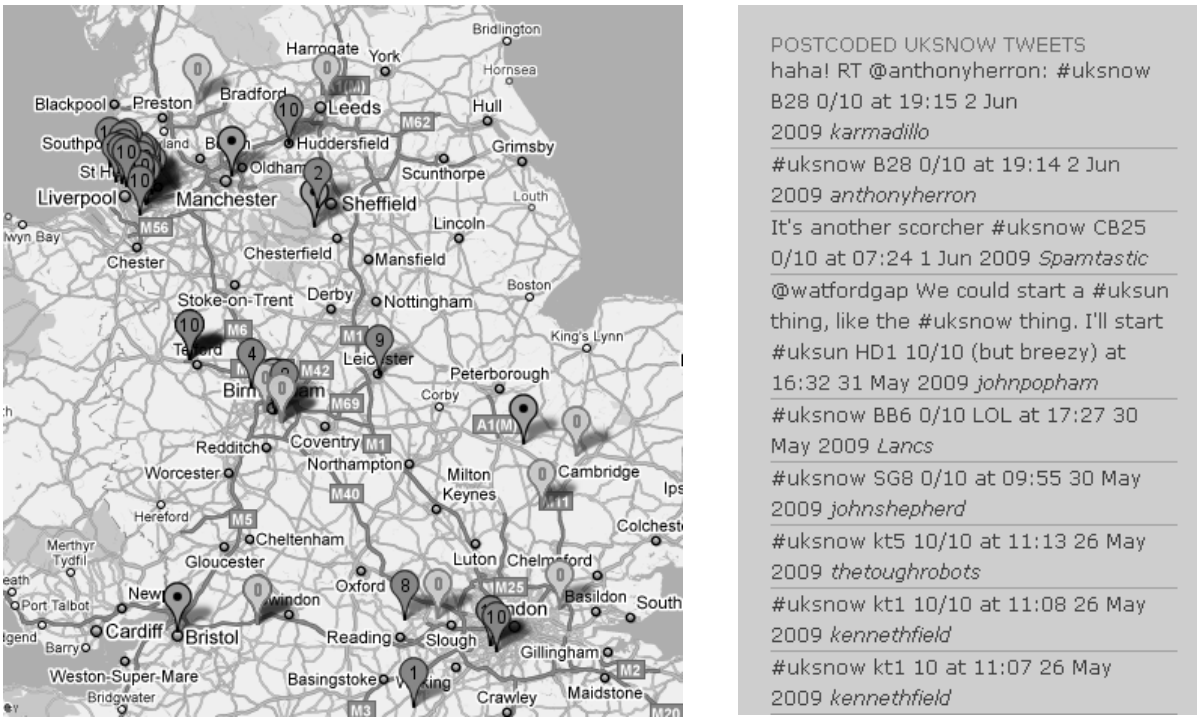



**Figure 1.** UK snow map mashup (Darbyshire, 2009)

The map is a good example of some of the cartographic problems of displaying data captured at a large scale based on an individual's location that is then translated to a small scale map product. Overlapping symbols hide a great amount of detail and rather that the reader inferring that a cluster of symbols is equivalent to a large snowfall, it simply reflects a large number of tweets being sent from that vicinity. The use of a standard Google base map creates unnecessary clutter – a problem with this style of mashup in general since there is no way to modify the base map upon which other detail is placed. Additionally, the areas with no markers do not necessarily reflect areas of no snow but areas of no tweets; a fact further illustrated by the urban bias in the information.

As this example demonstrates, there are useful spatial contexts that can be derived from tweets. However, there remain problems with the organisation and visualization of the mapped outputs. For instance, there is very little attempt to order information using themes, topics, a numeric measure or even place (other than by proximity) or make use of the information. Neither is there an attempt to address the cartographic problem of how to represent tweets sent from coincident or similar locations. Most location-based projects use the account-level location field but anything can be written by the user leaving it not very dependable. Finally, many tweets form part of a conversational timeline which means the temporal aspect of the visualization is not fully reconciled.

## 3. Method

### 3.1 Supporting collaborative investigation using Twitter map mashups

This proof-of-concept was originally developed to support student fieldwork in Malta and to demonstrate the value of combining social networking tools with map mashups for asynchronous collaborative learning. The #malta09 TweetMap made use of the Twitter search API to find the #malta09 (Kingston 2009 Malta field course) hashtag which was being used to denote content related to the fieldwork activities. The map illustrates tweets spatially arranged using GPS coordinates supplied within the tweet that took the following form:

#malta09 [latitude][longitude] [rating] text [twitpic URL]

e.g.

#malta09 35.8079 14.0425 3/10 mostly scrubland with some agricultural use http://twitpic.com/aabc

The latitude and longitude are used to position a marker on a Google Map (using the Google Maps API) and the rating is used to scale the size of the icon can be modified to allow rapid visualisation of student response to various questions set. The latitude and longitude values are passed to the standard Google Maps API function call GMarker along with the appropriately sized icon and added to the map. The content of the tweet is also copied into an HTML popup attached to the marker to allow for mouseOver and mouseClick events on the marker.

Figure 2 illustrates the map which focussed more heavily on location over representation. The justification for this was that students would be moving around the island of Malta for a week and would be making tweets referencing different spatial locations.

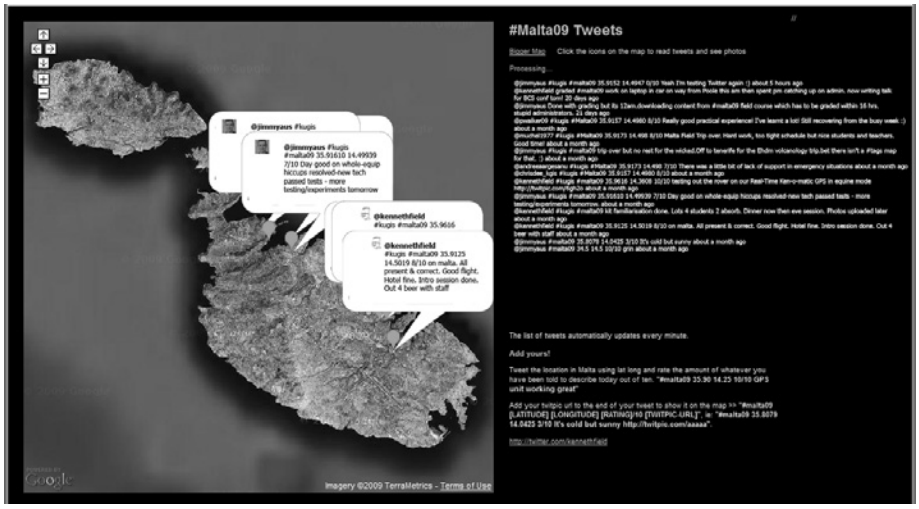



**Figure 2.** Twitter as a collaborative mapping tool on Malta

The Twitter map allowed a number of useful opportunities and builds upon Delphi principles for supporting asynchronous discussion (Dalkey 1969) and more recent implementations such as e-Delphi (Pike et al. 2009). Delphi is a technique used to elicit input from diverse groups with a goal to not necessarily reach consensus, but to identify key elements of a problem or points of agreement and disagreement. e-Delphi extends this web-based systems. Firstly, student groups were working in disparate locations and the map allowed a common framework in which students could post messages, work in collaboration with other groups and communicate with staff. Secondly, a number of tests were set for students that could only be successfully completed through collaboration, such as the identification of different land use zones and their physical appearance in different parts of the study area. Being able to take pictures that are uploaded to Twitpic gave students a mechanism to disseminate details of their case study area that assisted in the development of common frameworks for understanding and for data collection with the goal of creating a uniform land use classification and map. Thirdly, the map provided a community diary of each day's activities which provided scope for later discussion. Fourthly, in a practical sense the approach reduced the amount of post-processing in map creation. Finally, as a tool for staff, the map gave an immediate way of assessing where student groups were across the island. In a theoretical sense, this exercise enabled students to use the #malta09 TweetMap to develop a rudimentary ontology of landuse as well as an information product. The product (land use map) is a result of online interactions made possible through the mashup and which is dependent on the categories formed and the relationships between them.

### 3.2 Spatio-temporal design for Twitter map mashups

A second problem with many map mashups is the simple graphical representation that uses a single point of spatial fixity to represent location. For many maps this works well since locating a disparate set of phenomena reveals overall patterns. On the other hand, many phenomena occur in very close proximity which is often rendered as overlapping symbols, particularly at small scale zoom levels, on mashups. Further, mashups are beginning to reveal more than location and provide a way of visualising more complex information.

The AGI GeoCommunity 09 conference provided a second case study to explore ways of deriving improved meaning from mashed up tweets. Since tweets were sent predominantly from the conference venue itself they were spatially coincident. The task here was to demonstrate how the

temporal dimension of tweets can be visualized more effectively and how the spatial component can be modified to make tweets visible.

The Twitter search API was used to identify #geocom in tweets. The webpage which contains the map uses Javascript (specifically JSON) calls to the Twitter API to search for the appropriate hashtag. The API returns a formatted array of strings containing the tweets. The AGI map ignored the location of the attendees and focussed instead on the content of their tweets arranging them in different ways around the conference venue. The first approach used a random positional location (Figure 3) with more recent tweets denoted by larger place markers and common words being represented by similar colours. A pre-coded list of key terms was matched against comments in tweets and these allowed the symbols to be coloured in particular ways (e.g. green for 'keynote').

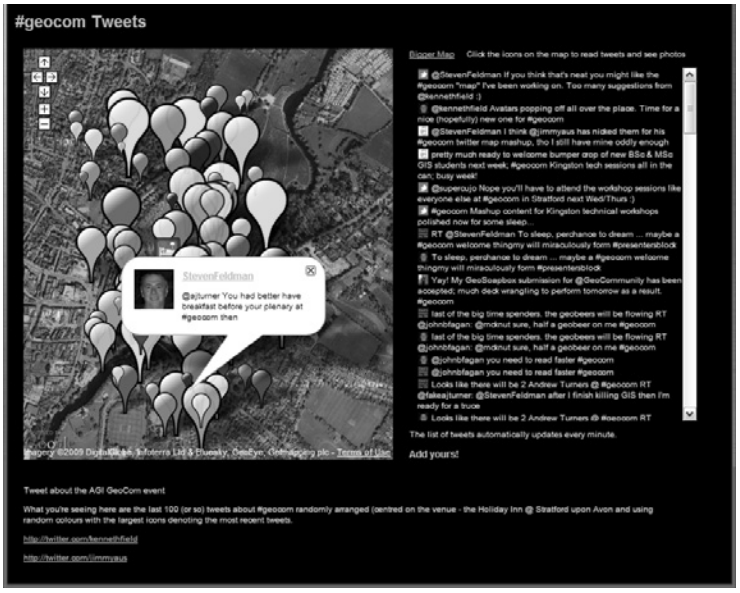



**Figure 3.** Random distribution of tweets around coincident location

The second approach used a series of cartographic representations for place markers arranged in a spiral around the venue with most recent tweets largest and on the outside of the spiral. The dimensions of the spiral were calculated within the Javascript

As with the Malta map, the content of the tweet is also copied into an HTML popup attached to the marker to allow for mouseOver and mouseClick events on the marker. Figure 4 illustrates a range of spatio-temporal visualisations of the tweets using this approach.

User testing revealed Figure 4(d) to be the most popular with 78% of the sample of 124 people surveyed at the conference saying it provided both a greater quantity and richer information. The original tweet is given a spatial context by proximity to the conference venue. It is further given a temporal context in a timeline of tweets all related to the conference. The use of avatars from Twitter profiles enabled viewers to readily see those who were contributing.

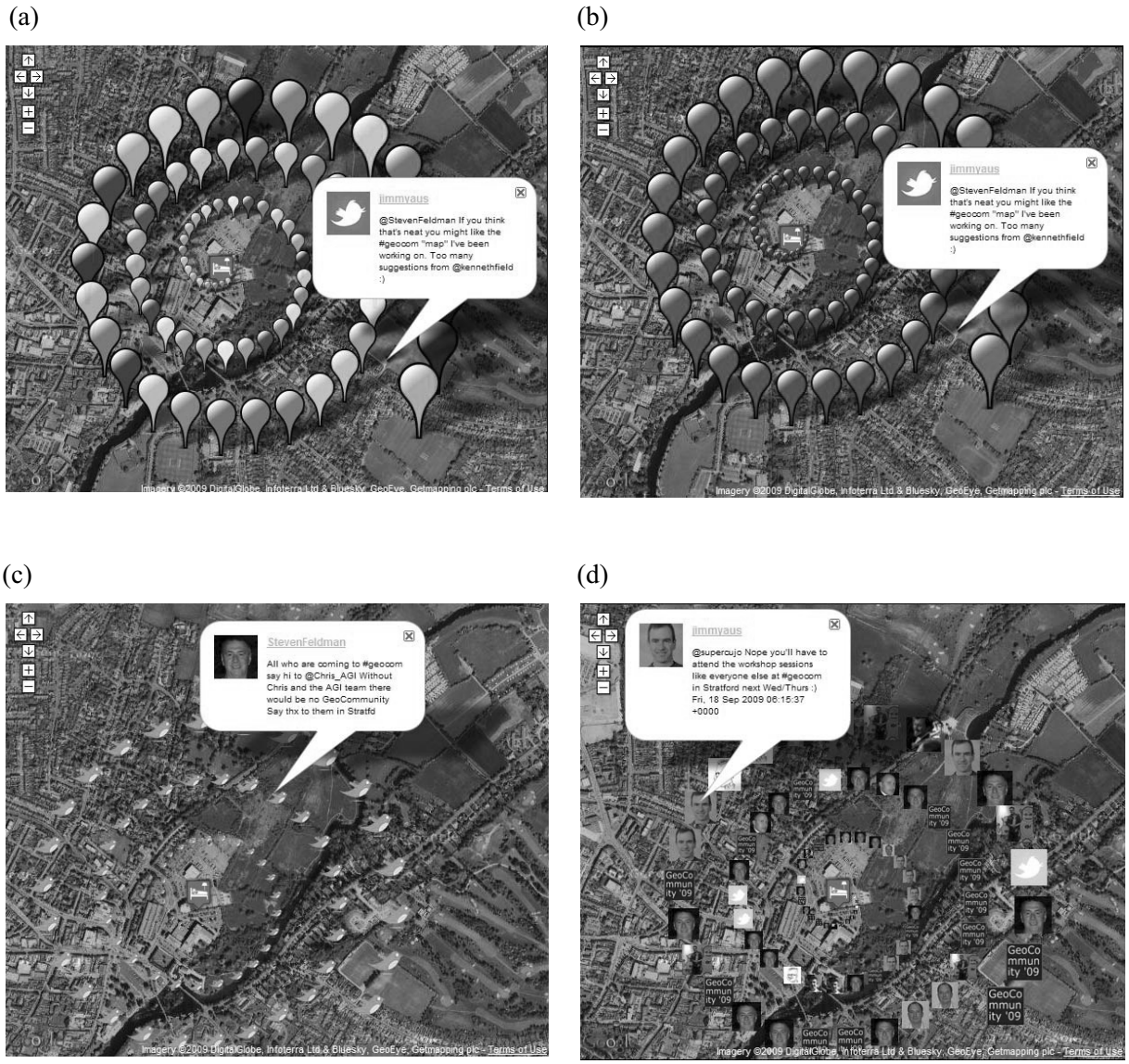



**Figure 4.** Spiral spatio-temporal timeline visualisations of tweets

## 4. Outcomes

Scope to create rich cartographic visualizations is apparent and provides scope for research in a number of ways in this rapidly expanding area of cartographic practice. The use of #hashtags and other key terms provides the ability to follow threads in tweets by individuals or groups but lacking spatial context these remain largely list-like. Extracting the spatial dimension reveals a contextual richness such as their location, their contribution or the place to which they are referring. The organization of this data in map form can reveal common organisational frameworks and support online collaborative learning. Innovative visualizations of tweets emanating from the same spatial location (e.g. a conference venue) illustrate approaches for effectively organizing information and to promote a richer, spatially organised micro-blogging environment.

## 5. Acknowledgements

This work is part funded by the transforming curriculum delivery through technology JISC e-learning programme Mobilising Remote Student Engagement (MoRSE), a collaborative project between researchers at Kingston University London and De Montfort University, Leicester.

## References

Clark, J. (2008) Twitarcs, (http://www.neoformix.com/Projects/TwitArcs/TwitArcs.html) accessed 16.2.10

Crampton, J. (2009) Cartography: maps 2.0, Progress in Human Geography **33(1)** pp.91-100

Dalkey, N. C. (1969) The Delphi Method: An Experimental Study on Group Opinion. The RAND Corporation

Darbyshire, K. (2009) UKsnow Twitter map (http://nocto.com/uksnow/) accessed 14.8.09

Field, K. S. (2009) Cartographic Twitterings, *The Cartographic Journal* **46(2)** pp59-61

Gibin, M., Singleton, A., Milton, R., Mateos, P. and Longley, P. (2008) An Exploratory Cartographic Visualisation of London through the Google Maps API, Applied Spatial Analysis and Policy **1(2)** pp85-97

Graham, M (2009) Neogeography and the palimpsests of place: Web 2.0 and the construction of a virtual earth, Tijdschrift voor economische en sociale geografie, **9999**

Haklay, M., Singleton, A. and Parker, C. (2008) Web mapping 2.0: The Neogeography of the Geoweb, Geography Compass 2(6) pp2011-2039

Marsh, B. (2009) #uksnow Tweets (http://benmarshcouk/snow/) accessed 16.2.10

Marsh, B. (2010) #uksnow Map 2.0 (http://uksnowbenmarshcouk/) accessed 16.2.10

Miller, C. (2006) A beast in the field: The Google Maps Mashup as GIS/2, Cartographica 41(3) pp 187-199

Pike, W., MacEachren, A. and Yarnal, B. (2009) Infrastructure for Collaboration. In: Sustainable Communities on a Sustainable Planet - The Human-Environment Regional Observatory Project edited by Yarnal B, Polsky C, and O'Brien J. Cambridge University Press: pp34-58

Schmidt, K. (2009) Social Collider (http://postspectacular.com/work/socialcollider) accessed 16.2.10

Wood, J., Dykes, J, Slingsby, A., Clarke, K. (2007) Interactive Visual Exploration of a Large Spatio-Temporal Dataset: Reflections on a Geovisualization Mashup, IEEE Transactions on Visualisation and Computer Graphics **13(6)** pp.1176-1183

## Biography

Dr Kenneth Field is Principal Lecturer in GIS and Course Director of undergraduate and postgraduate GIS courses at Kingston University. He is current Editor of The Cartographic Journal with particular interests in cartography, geovisualization and mobile GIS. He is currently in receipt of an Honorary Visiting SPLINT CETL Fellowship.

Dr James O'Brien is Principal Lecturer in GIS at Kingston University. His particular interests are in spatial databases, GIS software development, Semantic networks and ontologies and the application of mobile GIS for collaborative data capture.

# Land Evaluation Techniques Comparing fuzzy AHP with Ideal Point methods

Mukhtar Elaalem[1], Alexis Comber[2] and Pete Fisher[3]

[1]University of Leicester, Department of Geography Department, LE1 7RH

me84@le.ac.uk

[2]University of Leicester, Department of Geography Department

ajc36@le.ac.uk

[3]University of Leicester, Department of Geography Department

pff1@le.ac.uk

KEYWORDS: Land evaluation, Fuzzy Analytical Hierarchy Process, An Ideal Point, Uncertainty

## 1. Introduction

Land resources are gradually becoming scarce as increases in population place pressure on natural resources. The world population grows; an increase of food supply is urgently needed to meet those demands. In addition, Land use policy in developing countries frequently makes little use of technical information and when they do policy makers requires this to be interpreted into brief statements which eliminate technical details (Nwer, 2005). There are many approaches which are widely implemented in land evaluation such as: the USDA land capability classification (1961) and the FAO framework for land evaluation (1976). Some land evaluation techniques have been used in developing countries, but the information which utilized is often not linked to local knowledge (Clayton and Dent, 1993). Multi-criteria decision analysis methods (MCDM) such as analytical hierarchy process under fuzzy environment (FAHP) have employed with success in land evaluation technique (Parkash 2003). Very little studies have proven the use of MCDM methods to model land evaluation for agricultural crops. The main advantage with MCDM methods is that local knowledge can be taken into account to weight and select land characteristics that affect agricultural productions. The availability of GIS and MCDM methods allow combining knowledge derived form local knowledge to support land use planning and management (Malczewski, 1999). This paper compares land-use suitability analysis model for wheat using two MCDM methods; fuzzy AHP and Ideal Point methods.

## 2. Background of Multi-Criteria Decision Methods

MCDM approaches were developed in 1960s to assist decision-makers to incorporate many options, reflecting the opinions of the actors concerned, into a potential or retrospective framework. They were designed to define the relationship between the data input and the data output. MCDM can be separated into main two main groups of methods; multiobjective and multiattribute (Malczewski, 1999). In this paper, Fuzzy Analytical Hierarchy Process (FAHP) and Ideal Point methods have been selected to compare the model outputs for cash crops in study area. The AHP technique has the ability to incorporate different types of data and comparing two parameters at the same time by using the pairwise comparisons method; the base requirement for the AHP method (Saaty, 1977). An Ideal Point technique was selected to be used in this paper because it orders a number of alternatives on the base of their separation from the ideal Point and it employs a number of the distance metrics equations to produce the best alternatives ((Malczewski, 1999)

## 3. Methodology

There are many established techniques are extensively used for generating land-use suitability evaluation. The FAO framework with fuzzy AHP and Ideal Point method has been selected for a test area within part of Jeffara Plain in Libya. The main aim from this test is to incorporate local knowledge

from local experts and a literature review in order to the model of land-use suitability analysis. The paper methodology has been divided into four stages. These stages are**:**

### 3.1 Factors determining land-use suitability analysis for wheat

According to local experts, the study area is suitable for three cash crops such as barley, wheat and maize. For this paper, land-use suitability model for wheat has been developed using MCDM methods. A number of land characteristics that affect wheat production were identified after the discussion with the local experts for the study area selected. The main land characteristics affecting wheat production in the study area are: Soil texture, soil calcium carbonate, and rootable depth, available water holding capacity, soil organic matter, cation exchange capacity, soil salinity, soil alkalinity, soil drainage, and soil reaction, stones at surface, infiltration rate, erosion hazard and topographic characteristics.

### 3.2 Weighting parameters

Deriving weights for the selected map criteria (i.e. land characteristics map layers) is the base requirement for applying the fuzzy AHP and Ideal Point methods (Malczewski, 1999). Weighting factors for land- use suitability evaluation for wheat in this paper was obtained from local experts, through a pairwise comparisons statistical analysis in Idrisi environment. Four local experts in Libya have used their experience to generate weights for land characteristics wheat. One of four local experts result was accepted (table 1), because the Consistency Ratios (CR) was equal to 0.1 while, the rest were more than 0. CR $\leq$ 0.1 means that the comparisons of land characteristics were perfectly consistent, and the relative weights were appropriate for applying in land-use suitability evaluation models that use fuzzy AHP and Ideal Point methods. In addition to weighting factors, average weights for land characteristics for wheat were produced.

**Table 1**. Pairwise comparison matrix for wheat for the study area selected

| | % slop | soil texture | % $CaCO_3$ | % O.M | % ESP | AWHC | EC | % Stones | Soil drainage | Soil pH | CEC | Rootable | Infiltration | Soil erosion | average weights |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| % slop | 1 | | | | | | | | | | | | | | 0.032 |
| soil texture | 2 | 1 | | | | | | | | | | | | | 0.15 |
| % $CaCO_3$ | 2 | 1/3 | 1 | | | | | | | | | | | | 0.042 |
| % O. M | 2 | 1/3 | 1/3 | 1 | | | | | | | | | | | 0.035 |
| % ESP | 2 | 1/3 | 1/3 | 1/3 | 1 | | | | | | | | | | 0.028 |
| AWHC | 2 | 1/3 | 3 | 3 | 2 | 1 | | | | | | | | | 0.123 |
| EC | 2 | 1/3 | 3 | 5 | 2 | 1/3 | 1 | | | | | | | | 0.069 |
| % Stones | 2 | 1/3 | 3 | 2 | 2 | 1/3 | 1/3 | 1 | | | | | | | 0.043 |
| Soil drainage | 2 | 1/3 | 3 | 2 | 3 | 1/3 | 1/2 | 2 | 1 | | | | | | 0.051 |
| Soil pH | 2 | 1/3 | 3 | 2 | 3 | 2 | 3 | 3 | 2 | 1 | | | | | 0.132 |
| CEC | 2 | 1/3 | 3 | 2 | 3 | 1/3 | 1/2 | 2 | 2 | 1/3 | 1 | | | | 0.062 |
| Rootable depth | 2 | 1/3 | 3 | 2 | 3 | 1/3 | 2 | 2 | 2 | 1/3 | 3 | 1 | | | 0.08 |
| Infiltration rate | 2 | 1/3 | 3 | 2 | 3 | 1/3 | 2 | 2 | 2 | 1/3 | 1/3 | 1/3 | 1 | | 0.059 |
| Soil erosion | 2 | 1/3 | 3 | 2 | 3 | 1/3 | 2 | 2 | 2 | 1/3 | 3 | 3 | 3 | 1 | 0.094 |

### 3.3 Model structure

In this paper, the FAO framework for land evaluation with fuzzy AHP and Ideal Point approaches was employed to develop land-use suitability analysis for wheat. Fuzzy AHP and Ideal Point models are

given below:

### 3.3.1Land evaluation model using fuzzy AHP method

The fuzzy AHP procedure is considered one of the most common MCDM methods in resolving land suitability problem (Malczewski, 1999). The fuzzy AHP approach in this paper has been divided into five stages. These stages are summarized in figure (1).

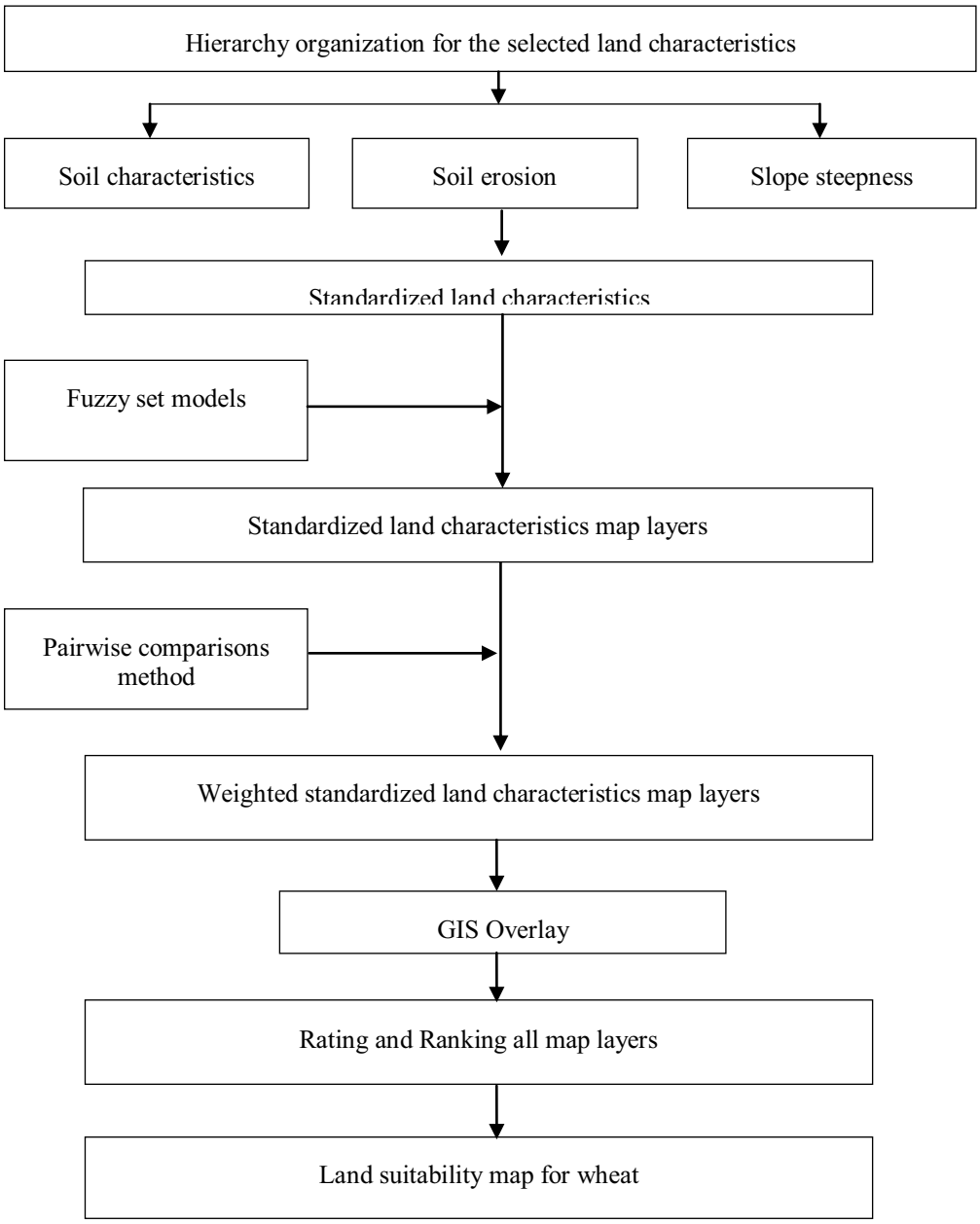



**Figure 1**. Fuzzy AHP method to the model of land-use suitability analysis for wheat

### 3.3.2Land evaluation model using an Ideal Point method

The fuzzy AHP procedure which described above has been extended in this paper to another five stages to be an Ideal Point land evaluation method. These stages are presented in figure 2.

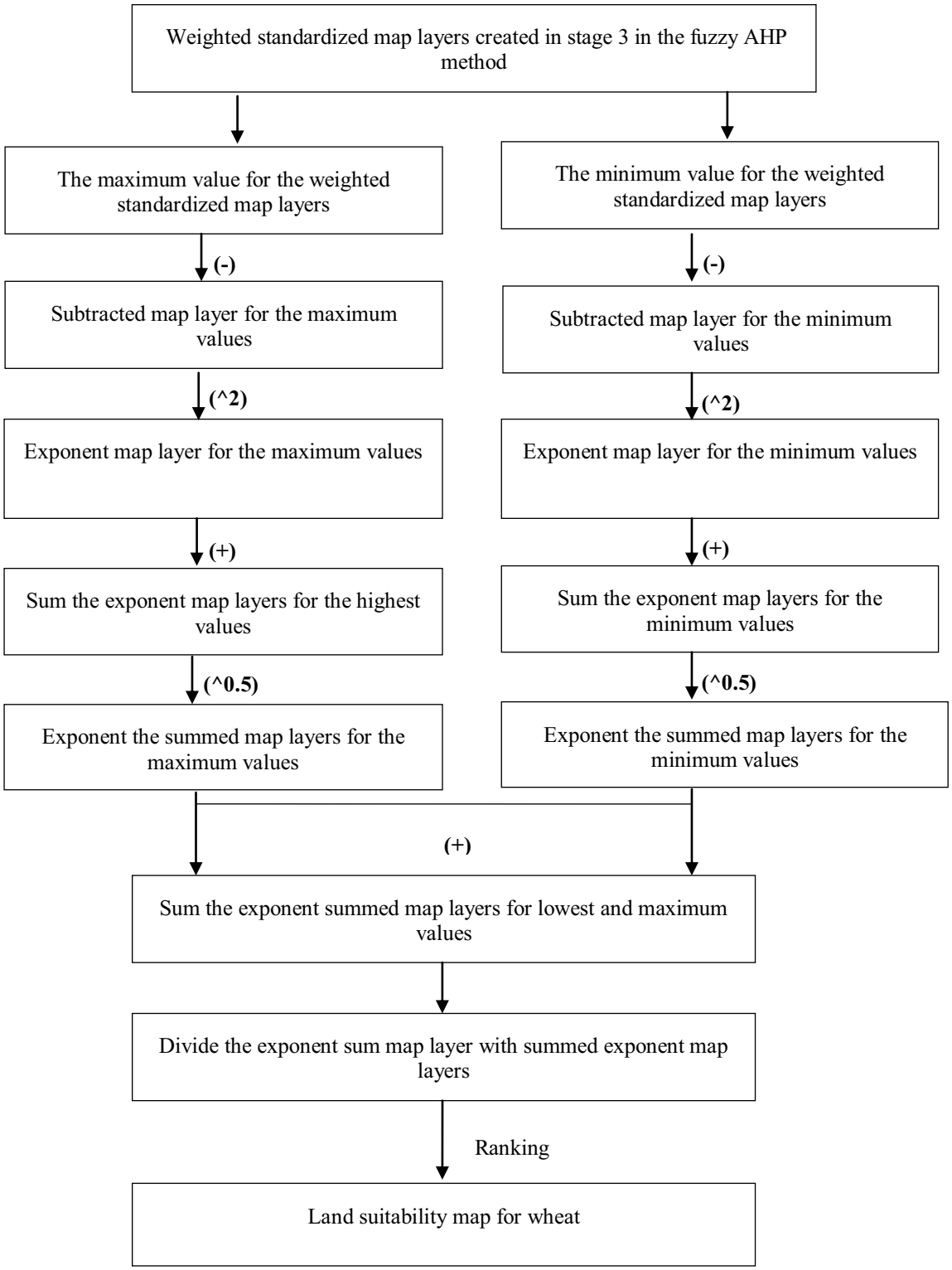



**Figure 2**. Ideal Point methods to the model of land-use suitability analysis for wheat

## 4. Results and Summary

The resulting maps for wheat from fuzzy analytical hierarchy process and Ideal Point methods were derived. Table 2 and 3 summarizes the results of suitability for wheat for fuzzy AHP and Ideal Point methods.

**Table 2.**The percentages of the overall membership function values derived from fuzzy AHP method

| MFs for wheat suitability classes | % |
|---|---|
| 0.79-0.63 | 1 |
| 0.63-0.49 | 19 |
| 0.49-0.36 | 69 |
| 0.36-0.29 | 6 |
| No data | 5 |

**Table 3.** The percentages of the overall membership function values derived from an Ideal Point method

| MFs for wheat suitability classes | % |
|---|---|
| 0.77-0.51 | 6 |
| 0.51-0.46 | 5 |
| 0.46-0.31 | 79 |
| 0.31-0.27 | 5 |
| No data | 5 |

With the fuzzy AHP and Ideal Point techniques it was very possible to obtain highly suitable and less suitable classes with parcel of lands have the highest and lowest MFs values. For this reason land parcel with high MFs values were ranked as classes 1 (most suitable classes) and parcels of lands with low MFs was ranked as classes 4 (less suitable classes). To determine the correspondence between land suitability maps they were cross- tabulated. The cross tabulation of fuzzy AHP and Ideal Point method for wheat is summarized in table 4. To evaluate the agreement between the model outputs, the kappa index values of agreement between the maps was calculated based on using different types of alternative hypothesis ($k \neq 0$, $k < 0$ and $k > 0$) and weights (none, liner and quadratic) (Table 5).From this table it can summarize that high agreement was obtained for land suitability maps, between fuzzy AHP and Ideal Point methods.

**Table 4**. Cross tabulation analysis between fuzzy AHP and Ideal Point methods

| | Ideal Point | | | | |
|---|---|---|---|---|---|
| Fuzzy AHP | 1 | 2 | 3 | 4 | Total |
| 1 | 2505 | 0 | 0 | 0 | 2505 |
| 2 | 112 | 31621 | 11781 | 0 | 43514 |
| 3 | 0 | 928 | 227773 | 0 | 228701 |
| 4 | 0 | 0 | 4776 | 14376 | 19152 |
| Total | 2617 | 32549 | 244330 | 14376 | 293872 |

**Table 5**. Shows kappa index values (fuzzy AHP vs. Ideal Point)

| Alternative hypothesis | Weights type | Kappa value | Kappa Interpretation |
|---|---|---|---|
| K≠ 0 | None | 0.82 | Almost perfect agreement |
| | Liner | 0.84 | |
| | Quadratic | 0.86 | |
| k< 0 | None | 0.82 | |
| | Liner | 0.84 | |
| | Quadratic | 0.86 | |
| K> 0 | None | 0.82 | |
| | Liner | 0.84 | |
| | Quadratic | 0.86 | |

## 5.Conclusion

Using different GIS functions to the model of land evaluation in the study area was one of the main objectives of this paper. Land evaluation model based on using fuzzy AHP and Ideal Point methods show that the percentages of land units which ranked as highly suitable and less suitable classes for wheat are very small and the MFs values for wheat which derived from the use the FAHP and Ideal Point had very little variation. This means the uncertainty was very less when the comparison between fuzzy AHP and Ideal Point methods to the model of land-use suitability analysis for wheat was made. From this paper a number of conclusions can be drawn. First, land characteristics affecting wheat production was very well organized and then assessed to fit into the framework of decision-making. Secondly, the use of decision-making methods to the model of land evaluation has facilitated the

incorporation of expert knowledge from different local experts. Thirdly, weighting of the selected land characteristics was made according to their relative importance with taken the crop requirement for wheat under local conditions into accounts.

**References**

Baja S D.M Chapman and D Dragovich (.2006). A Conceptual Model for Defining and Assessing Land Management Units Using a Fuzzy Modelling Approach in GIS Environment. Environmental Management . 29: 647–661.

Clayton, D and D. Dent. 1993. Surveys, Plans and People: A Review of Land Recourse Information and its Use in Developing Countries. Environmental Planning Issues 2, IIED

Davidson, D. A., S. P. Theocharopoulos, and R. J. Bloksma.1994. A land evaluation project in Greece using GIS and based on Boolean and fuzzy set methodologies. International Journal of Geographic Information Systems 8:369–384.

FAO (1976) A framework for land evaluation. Rome: Food and Agricultural Organization of the Unite Nations

Malczewski J (1999). GIS and Multicriteria Decision Analysis.USA and Canada, John Wiley & Sons

Ministry of Agriculture Fisheries and Food (1988). Revised guidelines and criteria for grading the quality of agricultural land.

Nwer, B., 2005. The application of land evaluation technique in the north-east of Libya. Published PhD thesis, Cranfield University at Silsoe.

Parkash T (2003) Land suitability analysis for agricultural crops: A fuzzy Multicriteria Decision Making Approach. Published Msc thesis.

Saaty, T. L., 1977. A scaling method for priorities in hierarchical structure: Journal of Mathematical Psychology 15.3, pp.34-39

USDA (1961) Land capability classification-Klingbiel A.A and Montgomery P.H .Soil conservation Agric Handbook No 210.USDA.Washinaton

# Developing a Statistical Methodology to Improve Classification and Mapping of Seabed Type from Deep Water Multi-Beam Echo Sounder (MBES) Data

Helen Caughey[1,4], Kazi Isthiak Ahmed[1,4], Paul Harris[1,4], Peter Hung[2,4], Urška Demšar[1,4], Sean McLoone[2,4], A Stewart Fotheringham[1,4], Xavier Monteys[3], Ronan O'Toole[3]

[1]National Centre for Geocomputation, National University of Ireland, Maynooth, Maynooth, Co. Kildare, Ireland
Tel. +353-1-7086455
[helen.m.caughey, kazi.ahmed.2009, paul.harris, urska.demsar, stewart.fotheringham]@nuim.ie
[2] Department of Electronic Engineering, National University of Ireland, Maynooth, Maynooth Co. Kildare, Ireland
[phung, sean.mcloone]@eeng.nuim.ie
[3]Geological Survey of Ireland, Beggars Bush, Haddington Road, Dublin 4, Ireland
[xavier.monteys, ronan.otoole]@gsi.ie
[4]StratAG – Strategic Research in Advanced Geotechnologies, http://www.stratag.ie

KEYWORDS: seabed type mapping, acoustic data, classification, spatial analysis, statistical analysis.

## 1. Introduction

This paper presents a recent collaboration between researchers of the StratAG group (Strategic Research in Advanced Geotechnologies, http://www.stratag.ie) and the Geological Survey of Ireland (GSI). The goal of the project is to develop a new statistical and spatial methodology to improve seabed type classification from deep water acoustic data.

GSI and the Irish Marine Institute (MI) have recently conducted extensive surveys of the Irish designated seabed area, under the Irish National Seabed Survey (1999-2005) and the INFOMAR programme (http://www.infomar.ie/, 2007-2013). This database presents a valuable resource for seabed mapping which is important for marine science research and a variety of applications, such as management of marine and coastal resources, fisheries, oil and gas exploration.

Collaboration between StratAG and GSI researchers focuses on seabed type mapping from data acquired with a Multi-Beam Echo Sounder (MBES). MBES backscatter datasets are massive as each collection line provides gigabytes of data. Therefore the classification process should be as automated as possible. It is also preferable to integrate datasets from different data acquisitions prior to classification. The project described in this paper looks at MBES data sets collected in Irish deep water areas. The work consists of two phases:

- Phase 1 uses spatial and statistical analysis to decide which datasets can be integrated and how – this work is on-going and is presented here;
- Phase 2 will consist of finding the best automatic classification methodology for integrated data and evaluation of classification quality against ground truth and/or existing seabed maps.

## 2. Seabed type classification from Multi-Beam Echo Sounder data

The Multi-Beam Echo Sounder (MBES) transmits and receives an array of acoustic beams across the axis of the ship. The intersection of the transmit pulse and the receive beams results in many simultaneous measurements across wide swath, with excellent resolution. Recent MBES systems use their larger angular width to record acoustic images providing the users with bathymetry and

reflectivity (backscatter) measurements at the same time (Augustin *et al*., 1994; Lurton, 2002; Mayer, 2006; McGonigle *et al*., 2009).

Seabed classification from multibeam sonar is achieved by the analysis of backscatter amplitudes. This is a complex process due to the diversity of ocean floor types and lateral inhomogeneity of sub-bottom layers (Xinghua and Yongqi, 2004, 2005; Arescon Ltd., 2001). Due to the large volumes of data acquired in MBES surveys, computer-assisted classification has become the logical choice to achieve statistically valid and objective segmentations (McGonigle *et al*., 2009; Hellequin, 1998; Hellequin *et al*., 2003, Cutter *et al*., 2003). Recently, focus has been given to classification based on the statistical nature of the image, irrespective of absolute calibration. One of the main commercial developments (specifically for MBES) based on this approach is Quester Tangent Corporation's Multiview (McGonigle, 2009; Preston *et al*, 2004; Preston *et al*., 2001). In this approach, statistical features are calculated from rectangular patches on the sea floor. Rectangular patches are distributed densely over the image to avoid many masked beams. The backscatter amplitudes within these rectangles are the raw materials for a set feature generation, called full feature vectors (FFVs) (Collins and Preston, 2002; Preston, 2009) and the result is a large matrix in which each column represents the values of one feature and each row contains all the features extracted from one rectangular patch. This matrix represents the FFV space, where each feature can be regarded as one new dimension/attribute.

## 3. Seabed type mapping from deep water MBES data – collaborative project between StratAG and GSI

### 3.1 Data description

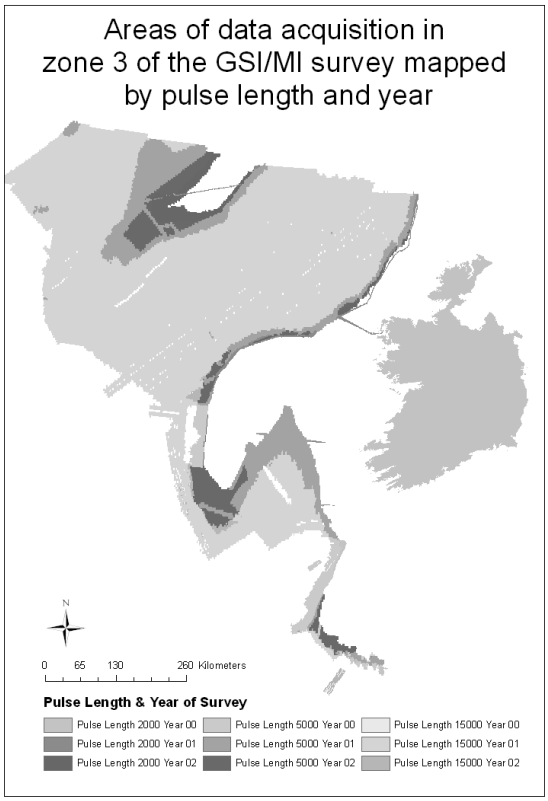



**Figure 1.** Map of data acquisition in zone 3 of the GSI/MI survey mapped by pulse length and year.

The survey area (Figure 1) extends between $57.4^0$N to $46.7^0$N and from $24.8^0$ W to $9.25^0$ W. The survey was conducted over a three year period (2000-2002). Three different pulse lengths were used

depending on the sea depth (2000ms, 5000ms and 15000ms). The majority of the survey was carried out using 15000ms pulse length in the year 2001. Each survey pass included five backscatter features (Q, P, C, M & S), where Q represents a quantile measure, P the 'pace' textural feature, C a 'contrast' feature, M the mean and S the standard deviation. The size of the feature-patch was pre-determined, as were the five features chosen for our analyses. 'Cross-length' backscatter (and bathymetry) data sets also exist and will be used to assess the accuracy of the classifications in phase 2.

### 3.2 Phase 1: data integration based on spatial and statistical analysis

The initial step in testing whether or not data integration was possible was to identify spatially overlapping data subsets. This should enable a comparative statistical investigation of distributions and relationships which would validate any subsequent joining of the parent data sets. The overlapping sections were obtained through ArcGIS functions, creating polygons around the centroid point files and extracting areas which overlapped between different years within the same pulse lengths. After an initial processing of the data it was decided that pulse length 2000 should be removed from the analysis. It only covered a small proportion of the survey area, consisted of hugely varied data and was considered unreliable in aspects concerning its means and consistency of acquisition. This left pulse lengths 5000 and 15000 from which 6 overlapping areas were identified over a spatially diverse area (Figure 2).


**Figure 2**. Areas of overlaps identified between the different years for pulse lengths 5000 and 15000. Names for overlaps (e.g. OL4, OL5…) are also given here as references from this point forward.

Each overlap consisted of multiple sub-areas of overlap not confined to one distinct spatial region. 'Pseudo Pairwise' overlapping data subsets were then identified automatically in the R statistical computing environment increasing the poor resolution of the overlaps obtained from ArcGIS. An example of this process for OL7 can be seen in Figure 3. This 'pairwise' data overlapped each other much better than before and was used in the subsequent analyses. For simplicity, we decided that we would only statistically assess similarity in the mean (M) feature values as this was deemed to be the most important feature for consideration.

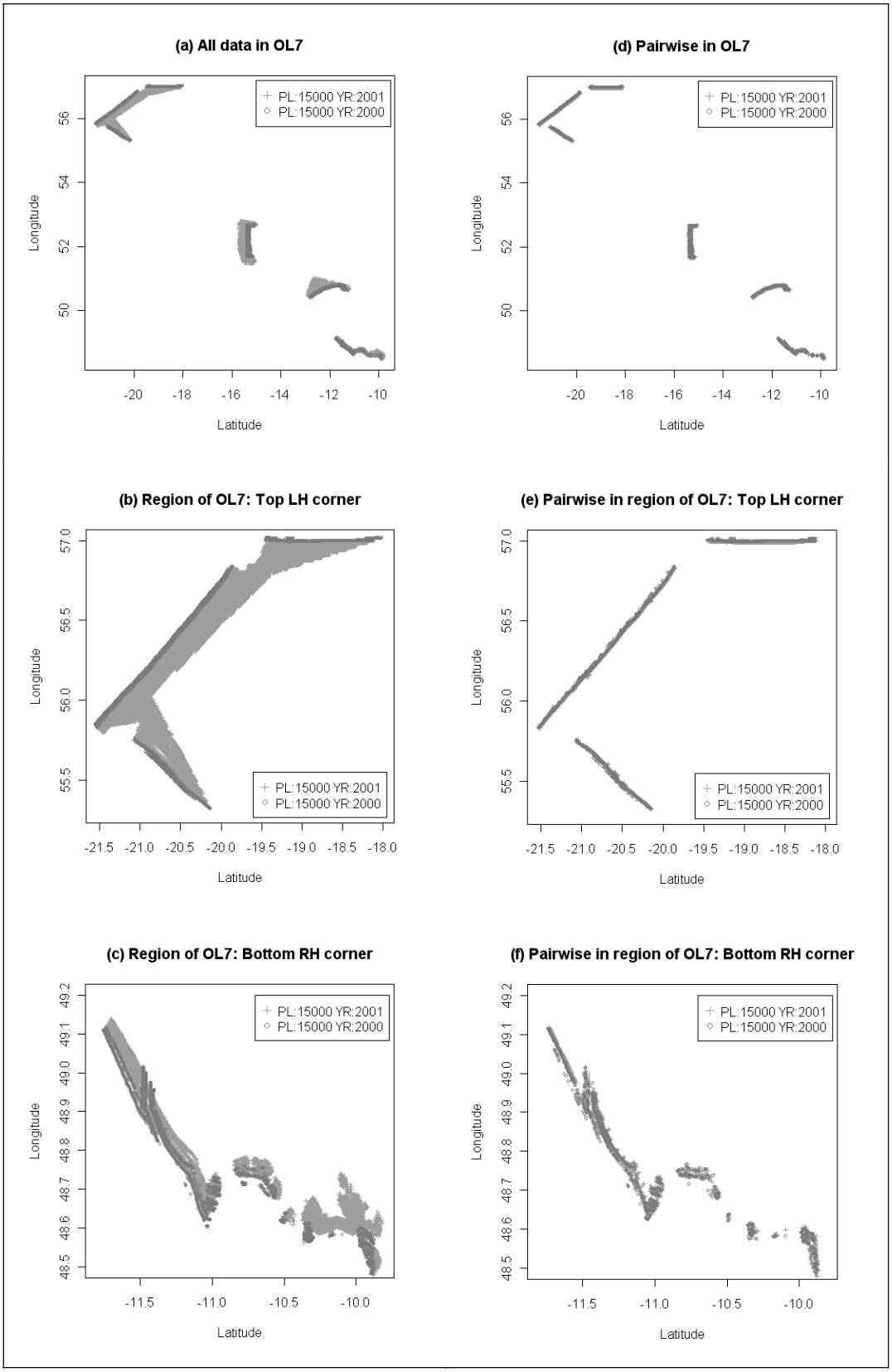



**Figure 3**. The computation of 'pseudo pairwise' datasets within the R statistical computing environment. Figures 3(a) – 3(c) show the overlapping datasets as they are calculated in ArcGIS while figures 3(d) – 3(f) shows the refinement of these within R.

A range of comparative analyses were carried out to assess the feasibility of joining the data. Here the distribution of the M feature across the six parent data sets was found to be broadly similar. Many overlapping datasets appeared unusual to their parent dataset and differences exist between each of the paired overlapping distributions. *F* tests and Kolmogorov-Smirnov (K-S) tests suggested variance and distributional differences for each overlapping data set pair. Variograms and scatterplots were constructed in an attempt to gauge autocorrelation and relationship differences. Using all this, most paired datasets showed some similarity, with the exception of OL8. However, only for OL6, OL7 and OL9 did the closest data pairs appear to exhibit the strongest scatterplot relationship in relation to data pairs that are far apart.

With this in mind, un-weighted and weighted (i.e. paired errors or differences that are close together are given more weight than paired errors that are further apart) mean errors, root mean squared errors and mean absolute errors were calculated. Here results suggested that parent data sets relating to OL4, OL5 and OL6 could possibly be joined. Weighting data pairs that are close together seemed only of value for OL5, OL8 and OL9 which differs to that perceived from the scatterplots.

Weighted correlation coefficients were also found where again overlapping data pairs are weighted according to its perceived importance. Based on Monte Carlo sampling 5% to 95% confidence intervals were computed where narrow intervals indicated confidence in the calculated correlation coefficient. Results suggested it would be unwise to join together parent datasets corresponding to OL4, OL5, OL8 and OL9. Robust correlations that are similarly weighted by the closeness in space of the data pairs were also examined. Here a box-car weighting scheme is specified and data pairs were removed in increments of 5% to a maximum of 60% of the paired overlapping dataset. The results of these robust *and* weighted correlations are seen in Figure 4 (with confidence intervals included). Evidence suggests it would be unwise to join together OL4, OL5 and OL8.

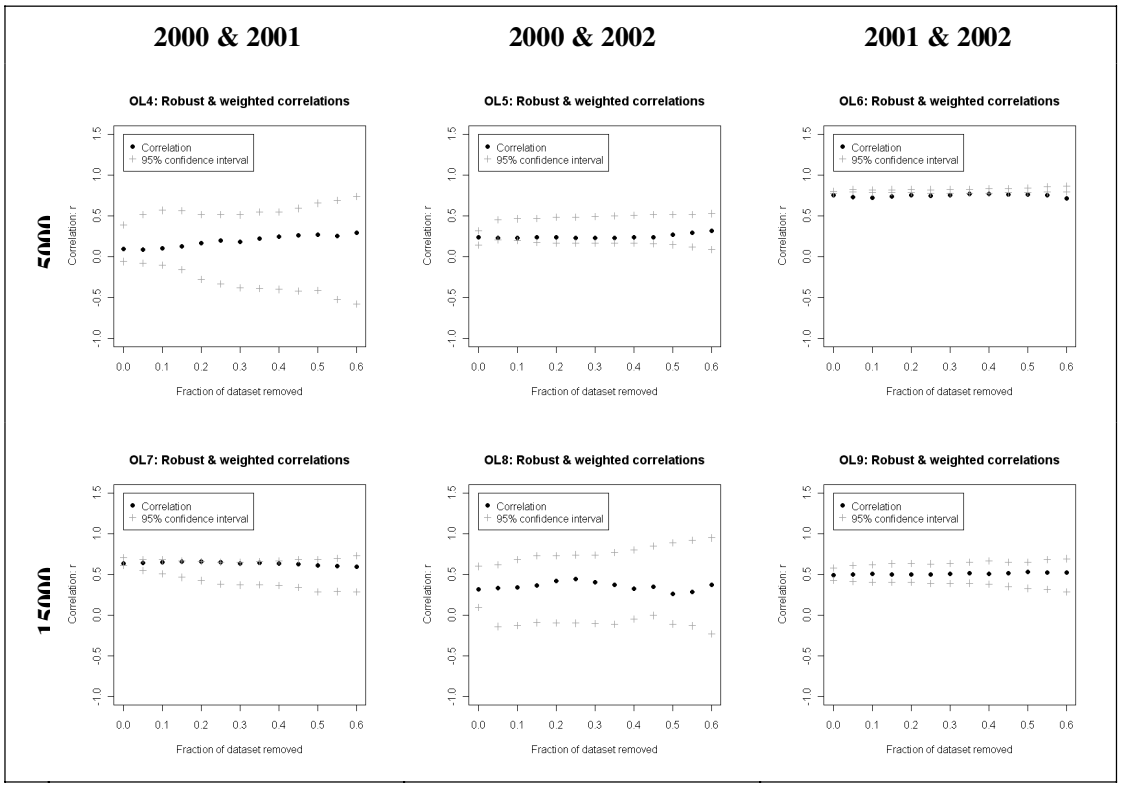



**Figure 4:** Robust and weighted correlations for paired overlapping M feature data.

The results of all the analyses are summarised in Table 1. Results suggest that it is not advisable to join OL8; PL15000/YR2000 with PL15000/YR2002. Results also suggest that it is unlikely that the following datasets should be joined together; OL4: PL5000/YR2000 with PL5000/YR2001, OL5: PL5000/YR2000 with PL5000/YR2002, OL7: PL15000/YR2000 with PL15000/YR2001. It is however likely that data sets OL6: PL5000/YR2001 with PL5000/YR2002 and OL9: PL15000/YR2001 with PL15000/YR2002 could be joined together.

| Overlap | Corresponding Datasets | Distribution | Hypothesis Tests | Spatial Autocorrelation. | Scatterplot. | Error statistics | Correlations |
|---|---|---|---|---|---|---|---|
| **OL4** | **PL5000/YR2000**<br>**PL5000/YR201** | No | No | | | | No |
| **OL5** | **PL5000/YR2000**<br>**PL5000/YR2002** | No | | No | | | No |
| **OL6** | **PL5000/YR2001**<br>**PL5000/YR2002** | | No | | | | |
| **OL7** | **PL15000/YR2000**<br>**PL15000/YR2001** | No | | | No | No | |
| **OL8** | **PL15000/YR2000**<br>**PL15000/YR2002** | No | No | No | No | No | No |
| **OL9** | **PL1500/YR2001**<br>**PL15000/YR2002** | | | | | No | No |

**Table 1**: Summary of results – Should parent data sets be joined together?

Currently we are working on focused scale investigations which may improve the accuracy of the given results. Analysis of the survey lines showed differences in the orientation of track lines that fell within the overlap areas. Therefore the same sets of analyses are performed on both geographically distinct overlaps and also where the lines are of the same orientation. Currently 30 separate focused scale overlap regions have been identified and are being investigated.

### 3.3 Phase 2: automatic classification of integrated data and evaluation of classification quality

Once the data sets are integrated, the next step will be automatic classification into acoustic classes. Although transformation of all data into a compatible form by merging is possible, difficulties arise in practice due to the influence of the physical properties of water columns and seabed types on the statistical features extracted, noise inherent in the MBES equipment and the relatively large area coverage of MBES in deep water applications compared to shallow or coastal water.

As the statistical analysis tends to suggest that data cannot be combined into one mega set for the purpose of classification, the focus going forward will be developing techniques for combining the classification results from different data sets. Initially, principal component analysis (PCA) and k-means clustering will be attempted. Other non-linear methods such as neural networks and quality threshold (QT) clustering (Heyer *et al.*, 1999) may be considered. The geological features (rocks, sediments, etc) corresponding to each labelled class will be determined using expert knowledge and comparison to results from other surveys. Validation of classification results will involve comparing the labels to classifications obtained from other physical sensing methods.

## 4. Conclusions

In this paper we present the first phase of a collaboration between StratAG and GSI on improving classification methodology for deep sea MBES backscatter data from an Irish designated seabed area. The goal of this phase was to use spatial and statistical analysis to decide which of the several data sets collected at different pulse lengths and at different moments in time were sufficiently similar to each other to be integrated into larger data sets. These larger data sets will be used in future seabed classification with the final goal to produce new improved seabed type maps for the entire deep water zone of the Irish designated seabed area.

The combined methodology of data integration and classification will be evaluated against ground truth information or existing seabed maps. Should the new classification method for deep water MBES backscatter data withstand these tests, it could be used by GSI (and other national marine agencies) to increase the quality of current seabed type maps.

## 5. Acknowledgements

Research presented in this paper is supported by a Strategic Research Cluster grant (07/SRC/I1168) – StratAG, awarded to the National Centre of Geocomputation by Science Foundation Ireland under the National Development Plan. This project is a part of the collaboration of StratAG with the Geological Survey of Ireland (GSI), who generously provided the MBES data.

## References

Arescon Ltd. (2001) *An Approach to seabed classification from multi-beam bathymetric sonar data* (http://arescon.com/pics/mbeam.pdf), last accessed on 7th May 2009.

Augustin JM, Edy C, Savoye B and Le Drezen E (1994) Sonar mosaic computation from multibeam echo sounder *OCEANS '94. 'Oceans Engineering for Today's Technology and Tomorrow's Preservation'. Proceeding* **2** pp433-438.

Collins WT and Preston JM (2002) Multibeam seabed classification *International Ocean Systems* **6(4)** pp12-15.

Cullen S (2003) Irish National Seabed Survey - An Introductory Overview *The Hydrographic Journal* **109** pp. 22-25.

Cutter JGR, Rzhanov Y, and Mayer LA (2003) Automated segmentation of seafloor bathymetry from multibeam echosounder data using local Fourier histogram texture features *Journal of Experimental Marine Biology and Ecology* **285-286** pp355-370.

Hellequin L (1998) Statistical characterization of multibeam echosounder data *Oceans'98* **1** pp228-233.

Hellequin L, Boucher JM, and Lurton X (2003) Processing of high-frequency multibeam echosounder data for seafloor characterization *IEEE Journal of Oceanographic Engineering* **28** pp78-89.

Heyer LJ, Kruglyak S & Yooseph S (1999) Exploring Expression Data: Identification and Analysis of Coexpressed Genes *Genome Research* **9(11)** pp1106-1115.

Lurton X (2002). *An Introduction to Underwater Acoustics: Principles and Applications*. Springer Verlag, Berlin-Heidelberg.

Mayer LA (2006) Frontiers in seafloor mapping and visualization *Marine Geophysical Researches* **27** pp7-17.

McGonigle C, Brown C, Quinn R and Grabowski J (2009) Evaluation of image-based multibeam sonar backscatter classification for benthic habitat discrimination and mapping at Staton Banks, UK . *Estuarine, Coastal and Shelf Science* **81** pp423-437.

Preston JM, Christney AC, Bloomer SF, and Beaudet IL (2001) Seabed Classification of multibeam sonar Images *Ocean's 01, MTS/IEEE conference 2001, Honolulu, USA* **4** pp2616-2623.

Preston JM, Christney AC, Collins WT, and Bloomer S (2004) Automated acoustic classification of sidescan images *Oceans'04, MTS/IEEE Techno-Ocean 2004, Kobe, Japan* **4** pp2060-2065.

Preston J (2009). Automated acoustic seabed classification of multibeam images of Stanton Banks *Applied Acoustics* **70(10)** pp1277-1287.

Xinghua Z and Yongqi C (2004) Seafloor sediment classification based on multibeam sonar data *Geospatial Information Science* **7(4)** pp290-296.

Xinghua Z and Yongqi C (2005) Seafloor classification of multibeam sonar data using neural network approach. *Marine Geodesy* **28** pp201-206.

## Biography

***Helen Caughey*** *and* ***Kazi Ishtiak Ahmed*** *are PhD students at the National Centre for Geocomputation (NCG) at the National University of Ireland, Maynooth. Helen's research is in Geophysics and Remote Sensing. She has a BA (Hons) in Media Studies and Geography and a postgraduate diploma in GIS and Remote Sensing both from NUI Maynooth. Ishtiak is working on Visual Analytics for large dataset with concentration on large scale acoustic seabed survey data. He has an MSc in Geoinformatics and Geodesy from the Royal Institute of Technology, Stockholm, Sweden and a B.Sc. in Civil Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. Dr* ***Paul Harris*** *and Dr* ***Peter Hung*** *are both Postdoctoral Researchers at the National University of Ireland Maynooth. Paul, an expert in spatial statistics, is based at NCG and has a PhD in Geostatistics from Newcastle University. Peter is at the Department of Electronic Engineering and has a PhD in electrical and electronic engineering from Queen's University Belfast, Northern Ireland. His research focuses on system identification, pattern classification and machine learning. Dr* ***Urška Demšar*** *is Lecturer in Geocomputation at NCG. Her research interests include geovisual analytics and mathematical analysis of spatial data. Dr* ***Sean McLoone*** *is Head of Department of the Department of Electronic Engineering, NUIM. His research interests lie in the general area of data based modelling and analysis of dynamical systems. Prof* ***A Stewart Fotheringham*** *is Director of NCG and Head of the StratAG group. His research interests are in spatial statistics, geographic information science and spatial modelling.* ***Xavier Monteys*** *is Marine Geologist at the Geological Survey of Ireland (GSI). He works on methodologies for seabed classification from different types of remotely sensed underwater data (acoustic, geophysical, seismic).* ***Ronan O'Toole*** *is a Marine Geoscientist at GSI.*

# Identifying and Testing Discontinuities in Surface Fitting Techniques

Chris Brunsdon
Department of Geography, Bennett Building, University of Leicester , Leicester LE1 7RH
Tel: +44 116 252 3843 | Email: cb179@le.ac.uk

KEYWORDS: discontinuity, surface, visualisation, house price, smoothing

## 1. Introduction

Surface-based approaches have frequently been used to analyse social and economic data. Using approaches such as kernel regression (considered here), or Loess smoothing, it has been possible to fit continuous surfaces to spatially reference social and economic data, such as house prices. The technique has often proved a useful tool in identifying trends in the data - for example one can identify areas of town in which housing is generally more costly.

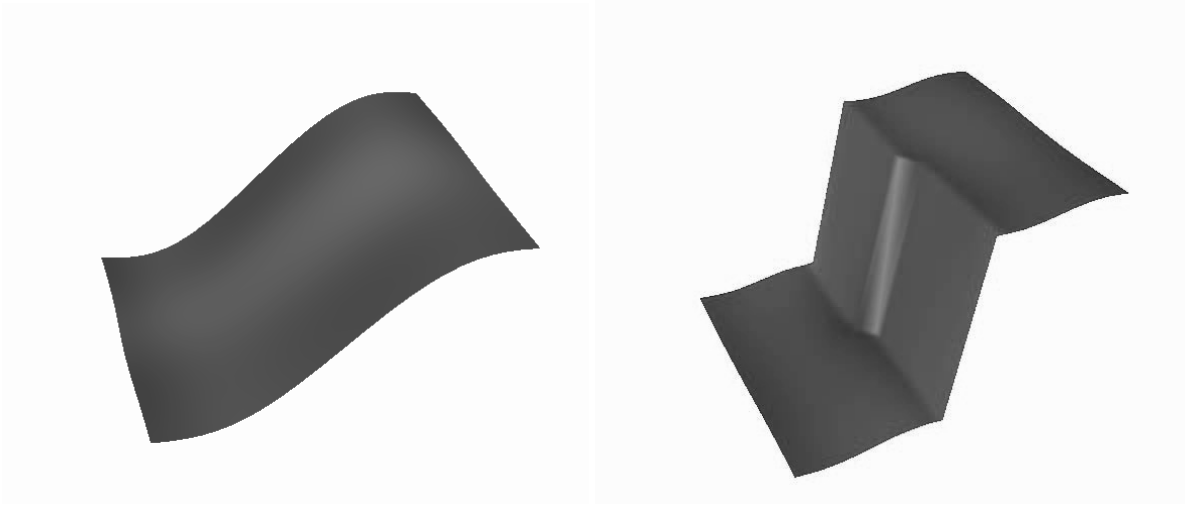
Thus, the idea of kernel regression is to estimate trend surfaces - so that if we have a set of points, say (*x,y*) and each one has a continuous scale attribute (say *z*) then the aim is to estimate the value of *z* at values of (*x,y*) other than those in the data set - essentially estimating a *z*-surface from a set of point observations of *z*. This is done by creating a *kernel* around the point (x,y) and taking a weighted mean of z-*values of* points in the vicinity of (x,y) - the weight decreasing the further the data points are away from (x,y). A typical kernel function might be

$$w = \exp\left(-\frac{d^2}{k^2}\right)$$

where d is the distance from the point at which we estimate *z* and a point in the data set, and *w* is the weight given to the *z* value associated with that data point. *k* is a smoothing parameter - the larger its value the smoother the trend surface.

For example this enables questions such as 'what price would we expect a house to be sold for in location (x,y)?' to be answered. This differs from *interpolation* in that interpolated values will always pass through the observed values of *z* whereas in a regression model this is not always the case. In a sense a regression surface passes smoothly through the *centre* of the observed z-values if viewed as a 3D point cloud - so observed and fitted values at points in the data set may differ. This is a sensible approach if the observed *z*-values may be subject to sampling variation or other random factors influence the sail price of a house.

A second difference - perhaps the key one here - is that not only smooth regression surfaces are considered, but also surfaces in which there may be discontinuities. An example highlighting the differences is shown below. In fact both surfaces were computed from the same set of point samples, but the RHS panel is the result of trend estimation with discontinuity detection, whereas the LHS panel uses a standard approach. Since the RHS approach applies smoothing over a window regardless of discontinuity, the effect is to smooth away this feature.



*Two surfaces: LHS without discontinuities, RHS with discontinuities.*

In this paper, two approaches to kernel smoothing - both modifying the basic kernel smoothing idea in different ways, will be outlined. The methods outlined are relevant to both physical and human geography data - in physical terms it allows terrain modeling with cliff edges, for example, but in this instance focus will be given to the fact that it can also be used to detect metaphorical fault lines in terms of social data - situations in which relatively affluent regions lie beside areas of high deprivation, or where house prices suddenly increase when a 'golden postcode' boundary is crossed. It is also worth noting that both methods *detect* discontinuities rather than work with locations deemed discontinous on an *a priori* basis.

The two methods will now be outlined:

**2. Anisotropic Smoothing**

Proposed by Perona and Malik (1990), this method operates on regular grid data. One way of smoothing grid data (without detecting discontinuities) is to replace the value of each pixel in the grid with a weighted average its immediate neighbours.

$$z_{ij}^{*} = \frac{w_{1,0}z_{i+1,j} + w_{-1,0}z_{i-1,j} + w_{0,-1}z_{i,j-1} + w_{0,1}z_{i,j+1}}{w_{1,0} + w_{-1,0} + w_{0,-1} + w_{0,1}}$$

The asterisk on the $z$ denotes an updated value, and this computation is applied to each $z$ in the 2D array. This only applies a small amount of smoothing - it is rather like a moving window approach where the window is just one pixel wide. An effect equivalent to a larger window can be achieved by repeatedly applying the operation just described.

Here the weights do not depend on the indices $i$ and $j$ - they are stationary - and the overall effect is similar to moving window smoothing. Unfortunately, for that reason, this approach also does not work well with discontinuities. However, suppose that the weights used were reduced in situations where the values of adjacent pixels were very different, or were on a rapidly changing part of the surface. This would reduce the 'smoothing off of edges' problem outlined in the last section.

One way to do this would be to make the weights depend on the slope estimates at each of the pixels - so that the influence of pixels on steeply sloping parts of the surface would be downgraded in the smoothing process. This could be achieved by a minor modification of the smoothing approach outlined above – instead of global weighting, for each pixel (i,j) set

$$w_{i,j} = f\left(s_{i,j}\right)$$

where

and $s_{i,j}$ is a slope estimate at pixel $(i,j)$. Typically, $f$ is a decreasing function, for example

$$\exp\left(-\frac{s_{i,j}^2}{h_2^2}\right)$$

so that the influence of pixels on a steep slope is downgraded.

As with the straightforward pixel based smoothing a single smooth operation takes place over a very tight window, but this time it does not over smooth when the surface changes rapidly. As before, the effect of using larger smoothing windows is achieved by repeated application. Typically, 20 or 30 applications are used, although this number can be chosen by cross-validation.

A final issue here is how the slope estimation is carried out. There are a number of possibilities - for example using Horn's method - however here, a fairly simple approach is used. This seems to be fairly effective. Assuming the grid spacing of the pixels is the same in the *x* and *y* directions a reasonable estimator of slope is:

$$s_{i,j} = \frac{(z_{i+1,j} - z_{i-1,j})^2 + (z_{i,j+1} - z_{i,j-1})^2}{4}$$

although other estimators such as Horn's method could be used.

Of course, this still only applies to gridded data, and in social and economic applications, irregular points are often found. The technique here requires the *z*-values to be known at the points where the trend is to be fitted, and using regular gridded data is a convenient way of ensuring this that will be the case. However, the question remains: "how can one apply this approach to irregular data?". The simple answer here is that some other method is first applied to the irregular data, resulting in a grid of values and subsequently the anisotropic diffisusion filter is applied to this grid. One issue here is that this involves the data undergoing *two* smoothing processes - and so there is a danger of oversmoothing. Usually this can be solved by carrying out the first stage with an interpolation approach – fitting the observed data values exactly – a naïve estimator setting each pixel to the value of the nearest observed point has been found to work well in practice. Next the anisotropic smoothing method is applied.

**3. An Alternative: Bilateral Smoothing**

This works on a slightly different principal, but still involves controlling the degree of smoothing when there is a large difference between the *z*-values of nearby points. In this case, one simply defines a kernel function in terms of *x,y* and *z* - typically kernel functions only depend on nearness in geographical space (x,y), but in this case nearness in attribute space is also considered. In this case the kernel function may by something like

$$w = \exp\left(-\frac{d_1^2}{2k_2^2}\right)\mathrm{I}(d_2 < k_2)$$

where d1 is a geographical distance as before and d2 is the absolute difference in *z* values between a pair of points. k1 and k2 are now both smoothing parameters - k1 functions as *k* does in a standard kernel regression model. For the term involving d2 and k2, *I* is an indicator function - it is zero if d2 exceeds k2 and one otherwise. The effect here is that standard kernel weighting occurs when two *z*-values are within d2 of one another, otherwise no weighting occurs. Thus if d2 is regarded as a significant enough difference to suggest a 'cliff edge' is between the points, no smoothing will occur. Since this method requires a z value at each place a prediction is made it is necessary to interpolate for values between known points after the smooth has been carried out.

**4. The Presentation**

In the proposed presentation these two smoothing methods will be introduced, and then illustrated with a number of examples using unemployment, house prices and Townsend scores of deprivation (a deprivation indicator) in Leicestershire, UK. These identify some potential 'faultlines' (Figure 1).

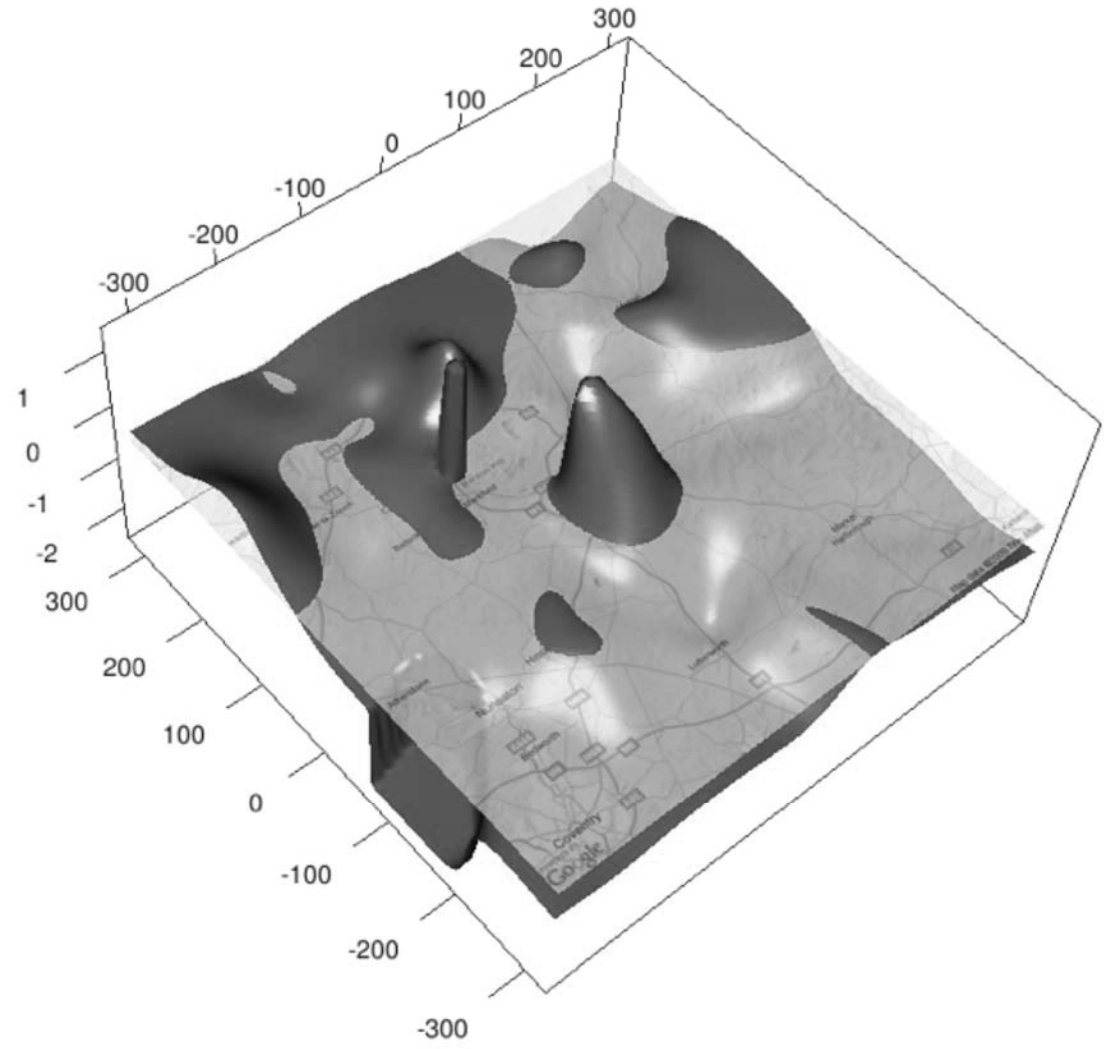



**Figure 1.** Surface of Unemployment Rate for Leicestershire with faultine detection. Note the 'spike' and sudden drop-off to the North of Leicestershire

Finally, to assess the reliability of these 'faultlines' a bootstrap analysis is carried out to see whether they could have occurred as an artifact of the methodology even in a situation where no sudden changes in the surfaces exist in reality. Shown in figure 2 are the upper and lower 95% envelopes for a house price surface for Leicestershire found using this approach. It may be seen that whereas both upper *and* lower envelopes are high for one 'plateau', this is not the case for some others – which may be due to outliers. Thus, this visualization method allows the plausibility of certain surfaces to be tested.

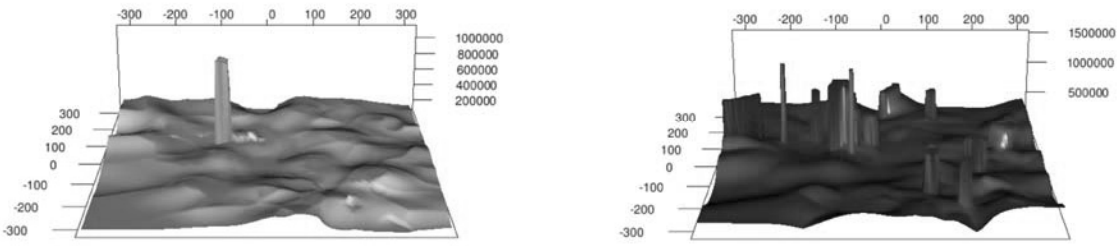



**Figure 2.** Upper and lower 95% envelopes for a house price surface for Leicestershire.

**References:**

P. Perona and J. Malik (1990). Scale-space and edge detection using anisotropic diffusion". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12 (7): 629–639.

C. Tomaso and R. Manduchi (1998). Bilateral Filtering for grey and Color Images, *Proceedings of the 1998 IEEE International Conference on Computer Vision*, Bombay, India

**Biography**

*Chris Brunsdon is currently professor of Geographical Information at the University of Leicester. He has interests in spatial analysis, spatial statistics and visualization techniques applied to geographical data.*

# GIS Based Spatial Modelling for Improving the Sustainability of Aggregate Mineral Supply in the UK

Chengchao Zuo[1], Mark Birkin[1], Graham Clarke[1], Fiona McEvoy[2], Andrew Bloodworth[2]

[1] School of Geography, University of Leeds, West Yorkshire, LS2 9JT
Tel. (+44 113 343 3330)
gy06cz@leeds.ac.uk
[2] British Geological Survey, Kingsley Dunham Centre, Keyworth, Nottingham, NG12 5GG

KEYWORDS: GIS, Spatial Interaction Model, Network Analysis, Aggregate, Carbon Footprint

## 1. Introduction

Aggregate minerals are essential in maintaining our economy and lifestyle, but their extraction, processing, and transportation causes constant concerns in regards to pollution and carbon emissions. Reducing this significant carbon footprint and environmental consequence in the face of accelerating demand for these construction materials is a major challenge facing the mining industry and its regulators over the next 30 years.

This research aims to provide a spatial decision support system for evaluating the sustainability of aggregates mineral supply and transportation options and to establish a framework for the evaluation of alternatives for the sustainable provision of aggregates in the future. This paper aims to describe the construction of a spatial decision support system to aid this process, which allows the user to investigate the environment consequence of specific policies such as demand forecasting, extraction licensing, and transport policies.

London will host the 2012 Summer Olympic Games, which has prompted a redevelopment of many of the areas of London. These projects require a large quantity of aggregates that sourced from all over the country. This development provides an ideal case study for the research; we will compare theoretical results from this research with benchmark data from real decisions regarding the supply of materials.

## 2. Methodology and Dataset

A GIS based spatial decision support system is being constructed, made up of four components: 1) Transport Model, which generates the Origin-Destination (O-D) cost matrix for the Spatial Interaction Model; 2) Spatial Interaction Model, which estimates the flows of aggregates between every pair of quarry and local authority district; 3) Environment Impact Assessment Module, which calculates the environment impact (e.g. carbon emission, noise etc.) from both producing and transporting aggregates; and 4) a Multi-Criteria Evaluation framework which evaluates the consequence (including both economic and environment considerations) of each policy. The flow chart below (Figure 1) illustrates the structure and data stream of the system:

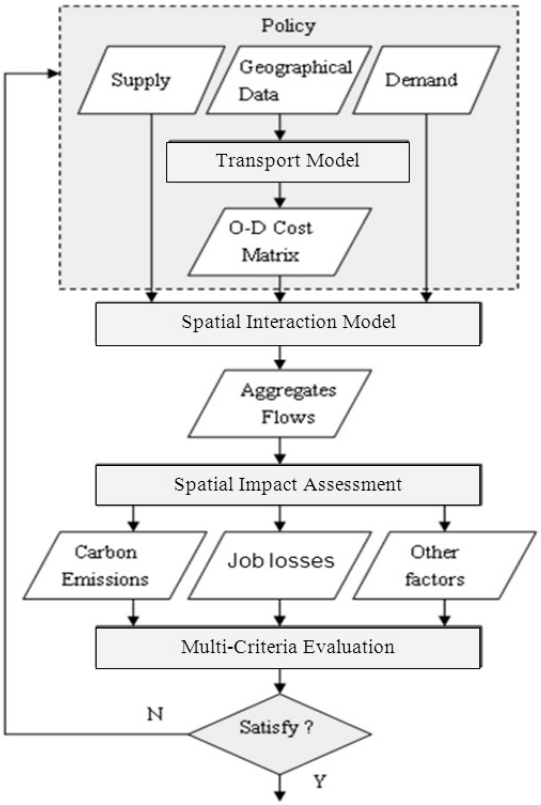



**Figure 1.** The Structure of Spatial Decision Support System

The Spatial Interaction Model (SIM) is a key element for the whole system as the interaction data is crucial for working through a spatial impact assessment. The SIM consists of four elements: a balancing factor K, two mass terms $O_i$ and $D_j$, and a travel cost function $f(d_{ij})$, so the basic equation can be written as:

$$\hat{m}_{ij} = KO_i D_j f(d_{ij}), \quad (i, j = 1,2,..., r; i \neq j) \tag{1}$$

where $\hat{m}_{ij}$ represents the volume of interaction from origin i to destination j. These models can be classified into 4 categories based on types the constraints used (Wilson, 1971): 1) unconstrained model, which means both supply and demand side are open to change, 2) supply constrained models, 3) demand constrained models, and 4) doubly constrained models, which means both supply and demand side are constrained to known totals.

For mineral planning purposes, the doubly constrained model is used. In this case, the constant (k) is replaced by two sets of balancing factors which ensure that the individual interactions sum to the known origin and destination totals for mineral supply. For practical applications, the cost component is also disaggregated into two competing elements, relating to costs of supply by road or railway. This arrangement is illustrated in Figure 2. Here we see an interesting policy problem is that increased supply by rail would help to increase the sustainability of aggregate mineral supply, but transhipment by rail is heavily constrained by both the capacity and linkages to the existing network.

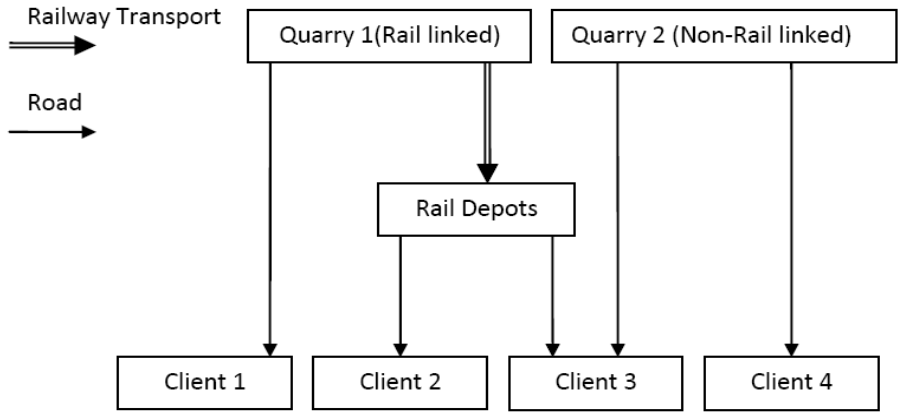



**Figure 2.** The Structure of Transport Model

Once the interaction matrix is obtained from the Spatial-Interaction model, the spatial impact factors and carbon emission can be evaluated through the spatial impact assessment module. Finally, Multi-Criteria Evaluation system will be built to evaluate the policy alternatives based on these environment impact factors, carbon emissions, plus economic benefits.

## 3. Scenario Settings and Modelling Results

The system developed in the preceding sections is now illustrated in an application which not only demonstrates the operation of the system, but also aims to explore the impact of variations in the policy settings on the environment consequence of the aggregate transportation.
A baseline model is built using the 2005 aggregate statistics data, which generates the interactions of aggregates at quarry-district level. The following table (Table 1) indicates the statistic fit of the model in contrast with the observed data (QPA, 2006):

**Table 1.** Comparison of Observed and Modelled data

| All Aggregates | *Observed* | *Modelled* |
|---|---|---|
| Market-share of Rail | 6.8% | 6.9% |
| Average Rail Delivery Distance | 163km | 181km |
| Average Road Delivery Distance | 38km | 45km |
| Tonne-km by rail transport | <2.7bn | 2.1bn |

Newham is the London borough where the Olympic Park will be mainly constructed. The pie chart below (Figure 3) indicates the proportion of each aggregate mineral used in this district in the year 2005 (BGS, 2007).

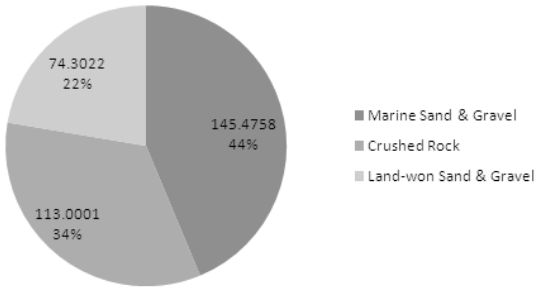



**Figure 3.** The Proportion of Different Aggregate Minerals Consumed in Newham

The map (Figure 4), which is generated by the spatial interaction model, explains the aggregates volumes flow to Newham by transport system in the year 2005.

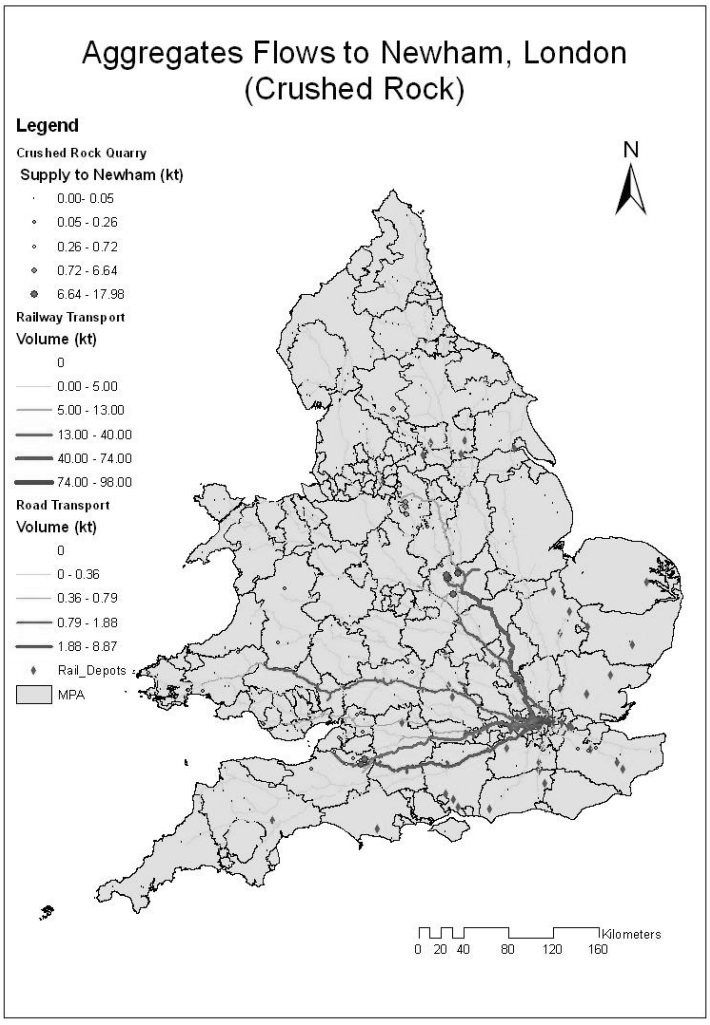



**Figure 4.** Aggregates Flows toward Newham

The blue lines represent the routes of aggregates by railway, and the purple lines represent the flows via road network. The volume of aggregates is indicated by the thickness of the line. The model

results shows that the crushed rocks consumed in Newham are mainly sourced from the Midland and Southwest area and mainly carried by the railway system.
Assuming in 2009, that 600 thousand more tonnes of aggregates will be consumed due to the construction of the London Olympic Park (BGS, 2008), the next chart shows the market share of aggregates by transport mode:

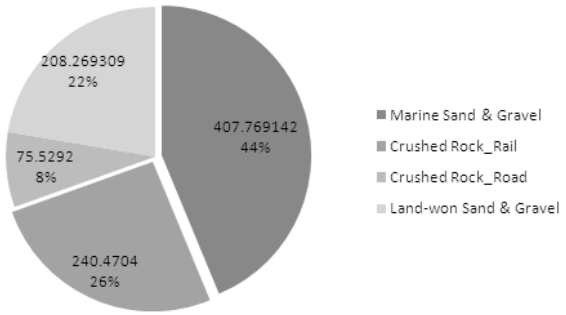



**Figure 5.** The Market-share of Different Transport Mode

where all land-won sand and gravels will be transported by road; approx.75.53kt of crushed rock will be delivered by road ( which take account 8% of all aggregates); 240.47kt of crushed rock are transported by railway; and for the marine source aggregates, all of them will be transported through waterway and then landed at the wharves on the river Thames.
According to the competitive nature of the rail and road transport modes in the model, if the relative cost of rail transport could be decreased, more railway delivered aggregates could be expected. Figure 6 shows the market share of different transport modes when the relative cost of rail transport is halved. This is to enable the proportion of 'sustainable' transported (through rail way or waterway) aggregates can achieve a 75% rate (44% of which are marine source aggregates which delivered by waterway, and 31% of them are rail-delivered crushed rock which was 26% in 2005).

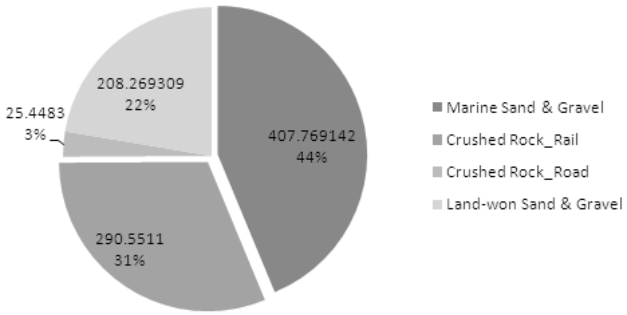



**Figure 6.** Proportion of different transport mode when relative cost halved

## 4. Conclusion and Discussion

This paper describes an ongoing project; so far, we have completed the transport model and spatial interaction model, and established a framework to evaluate different policy settings. The system

successfully simulated the interactions of aggregate minerals at the small area level and shows a good response to different policy scenarios.

For the next step, a more representative transport cost function can be introduced to replace the current distance based model, which should take time consumption, fuel cost, and capacity constraints into account. A spatial impact assessment model will be built to estimate the carbon emissions and other environmental or social-economical impact factors.

In future case study application, we will seek to illustrate a much more extensive range of policy scenarios to which the model may be applied. The analysis of Olympic Park illustrates possible uses of the model in the assessment of future spatial demand scenarios. This work could be extended to consider the impact of regional house-building strategies, or a more comprehensive analysis of future demographic change and its associated infrastructure requirements. Alternatively, specific issues such as the need for aggregate-intensive investments in things like flood protection in the face of global climate change might be explored. On the supply-side, scenarios such as the substitution of marine aggregates, and in the long-term the effect of aggregate recycling, might be considered. A third set of policies relating to the movement of aggregates can also be evaluated. Relevant options could include a restriction on the production or movement of minerals within the National Parks, or the resuscitation of inland waterways as a medium for transportation.

## 5. Acknowledgements

We are grateful to Leeds University Centre of Spatial Analysis and Policy, and British Geological Survey for their technical and funding supports.

## References

Arnold, P., Peeters, D., and Thomas, I., (2004), Modelling a rail/road intermodal transportation system. *Transportation Research Part E* **vol. 40**, pp.255-270

British Geological Survey (2007) Collation of the results of the 2005 Aggregate Minerals Survey for England and Wales, Department for Communities and Local Government, London

British Geological Survey (2008), The need for indigenous aggregates production in England, *British Geological Survey Open Report*, OR/08/026, Nottingham

Quarry Products Association (2006), *QPA Sustainable Development Report 2005 Update,* Grillingham House, London

Wilson A.G. (1971), A family of spatial interaction models, and associated development *Environment and Planning* **vol.3** pp.1-32

## Biography

***Chengchao Zuo*** *is a PhD student in the School of Geography, University of Leeds. His interests include GIS and spatial modelling. His PhD project aims to reduce the carbon footprint associated with primary aggregate mineral consumption in SE England.*

***Mark Birkin*** *is Professor of Spatial Analysis and Policy in the School of Geography, University of Leeds. His research interests include modelling societies and their behaviour, the use of information and communication technologies in the public, industrial and commercial sectors, and the development of Geographical Information Systems for human and environmental analysis. He is co-editor of the journal Applied Spatial Analysis and Policy.*

***Graham Clarke*** *is Professor of Human Geography in the School of Geography, University of Leeds. His research interests include GIS, urban services, retail and business geography, urban modelling and continuing professional education.*

***Fiona McEvoy*** *is a Principal Economic Geologist and former Team Leader for Minerals Policy Analysis at the British Geological Survey. Her interests include minerals policy development, in particular the protection of mineral resources, and the application GIS to mineral exploration. Fiona is a Professional Member of the Institute of Geologists, Ireland.*

***Andrew Bloodworth*** *is Head of Science for Minerals and Waste at the British Geological Survey. His own interests include spatial planning policy and regulatory issues associated with minerals extraction. He was formerly the Mining Adviser to the UK Department for International Development. Andrew is a Chartered Geologist and an Associate Member of the RTPI.*

# A Map Mashup for the Rural Urban Classification of England and Wales

Maurizio Gibin[1], John Shepherd[2]

[1]Department of Geography, Environment and Development Studies,
Birkbeck College, Malet Street, London
Email: m.gibin@bbk.ac.uk

[2]Department of Geography, Environment and Development Studies,
Birkbeck College, Malet Street, London
Email: j.shepherd@bbk.ac.uk

KEYWORDS: Rural Urban Classification, Map Mashups, Neogeography, Geovisualisation,

## 1. Introduction

Since the advent of Neogeography (Turner, 2006), new research has been dedicated to the visualisation of geographical and statistical data (Hudson-Smith et al., 2009) through web platforms. Map mashups are quite diffused Neogeography applications that combine different data sources together and provide base reference maps that add context and value to the visualisation of choropleth maps (Gibin et al., 2008). A number of academic projects are currently investigating the possibility of creating immersive web mashups that allow users to visualise, query, navigate and analyse Census data. (Gibin et al., 2009). This paper presents the work in progress of creating a map mashup for displaying the Official for National Statistics (ONS) Rural Urban Classification developed by the Rural Evidence Research Centre (RERC) based at Birkbeck College, University of London. First, the paper will briefly describe how the Rural/Urban Classification was created and then will introduce the main requirements for the future map mashup and will present the some initial tests.

## 2. The Official for National Statistics Rural Urban Definition

In 2003 a review of definitions of urban and rural Area used in England and Wales identified over 20 different ways of identifying these basic geographic categories for research and policy purposes (SERRL, 2002). Three types of indicator were generally employed either singly or in combination namely population size of settlement, land use and economic structure, including patterns of interaction. In addition, some form of population density measure appeared in most definitions , though it was rarely recognised that this measure is heavily scale dependent. The review noted that in the 'post-productivist' era the three core ideas behind urban-ness rurality had 'drifted apart' both functionally and geographically as exemplified by the decline of agricultural employment (by 2001 it accounted for less than 2 percent of total employment in England) and by the 'counter-urbanization of people and jobs out of the cities and into the countryside. Because of this spatial 'dislocation of key underlying ideas and rapid changes in economic structures as exemplified by counter-urbanization and loss of services the review recommended that a new definition should focus on the 'more enduring' aspects of rural areas namely, the morphology of settlement as expressed in urban land use measured via a range of geographic scales.

The new Rural Urban definition (Bibby and Shepherd, 2004) identifies two main features of the rural settlement landscape: settlement *morphology* and settlement *context*. First, every residential postal address in England and Wales (around 23 million items) is allocated to a 1 hectare grid. The next step is the calculation of a *density profile* for each cell in the grid using a constant numerator but varying denominators at 200m, 400m, 800m and 1600m around each cell. The method was found to be capable of identifying at least nine distinct rural settlement types including rural town, town fringe, village, village envelope, hamlet and isolated dwellings. The same principle was applied at wider geographic scales to

identify the population *context* in which a rural settlement type is found. This can be interpreted as the generalized accessibility of a settlement (i.e. average expected distances) to town shopping, local authority services and emergency services. In the case of the new rural definition generalized densities were calculated at three scales: 10,000m, 20,000m and 30,000m.

The final stage in the definition process is the classification of the basic national census areas (i.e. Census Output Areas, Super Output Areas, wards etc.) according the proportion of population in each settlement type within the census area concerned. In the case of the *settlement morphology* measure an area is classified according to the majority (>50%) of resident population in one of three settlement types: rural town, village and 'dispersed' dwellings. In the case of *settlement context* an area is 'scored' as 'sparse' if it has the majority of its population at a specific level of average density in all three geographic scales. This results in a two-level definition of 'rural' census areas as shown in Figure 1.

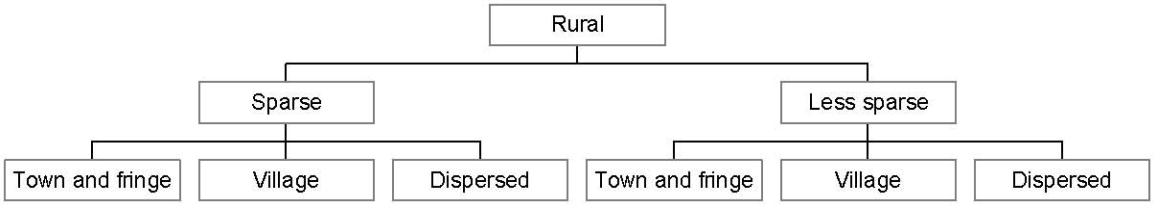



**Figure 1:** The Structure of the Rural Domain of England and Wales

**3. The RERC interactive maps**

The Rural Urban definition has been through three stages of formal consultation/evaluation. Although these cannot be expected to yield the full range of user issues they offer some valuable insights into the users experience. In addition, the Rural Evidence Research Centre has fielded detailed, place based, questions on the use of the definition and has harvested users feedbacks about the interactive maps available at the RERC website (Figure 2). The main conceptual concerns (or, in some cases, misinterpretations) of the definition revolve around its singular focus on built over land as represented by the density and clustering of grid squares populated by residential delivery points. When translated into rates of occurrence of settlement classified grid squares within census units the difficulties are that (a) this subjugates any notion of 'open country' which is, for most people *the* distinguishing characteristic of 'rurality', (b) that the classification of census areas into settlement types is done via an arbitrary '50 percent rule' which may, if changed' lead to a quite different classification and (c) that in some cases this majority may arise from the aggregation of classified grid squares that are of similar settlement type but spatially separated.

The question is, then, how can we enable the user of the rural urban definition, particularly at the very local level, to 'evaluate individual areas' and provide evidence for the consideration of the 'other issues' that enter into a valid and shared interpretation of the definition. Beneath the complex, multi-level, multi shaped 'grid' of census and other administrative areas, therefore, is a settlement pattern expressed as 'built-upness' ranging from individual isolated dwellings, through hamlets to villages and small rural towns. Interpretation of the significance of the rural urban definition for a given, census or other statistic rests, therefore, on understanding the territorially unique relationship among spatially averaged grids representing settlement morphology, the 'real world' pattern of settlement and the spatially averaged statistics reported for the defined area (Figure 3). The current interactive map visualisation (Figure 2) hosted at the RERC website is not able to solve the issues mentioned above, for two main reasons. The Graphic User Interface (GUI) does not ease the creation of maps and complicate the zooming and panning experience by a series of buttons and drop down lists (Haklay and Zafiri, 2008). The second reason is the lack of base reference maps that could allow the user to contextualize the choropleth maps especially at very local level (Figure 3).

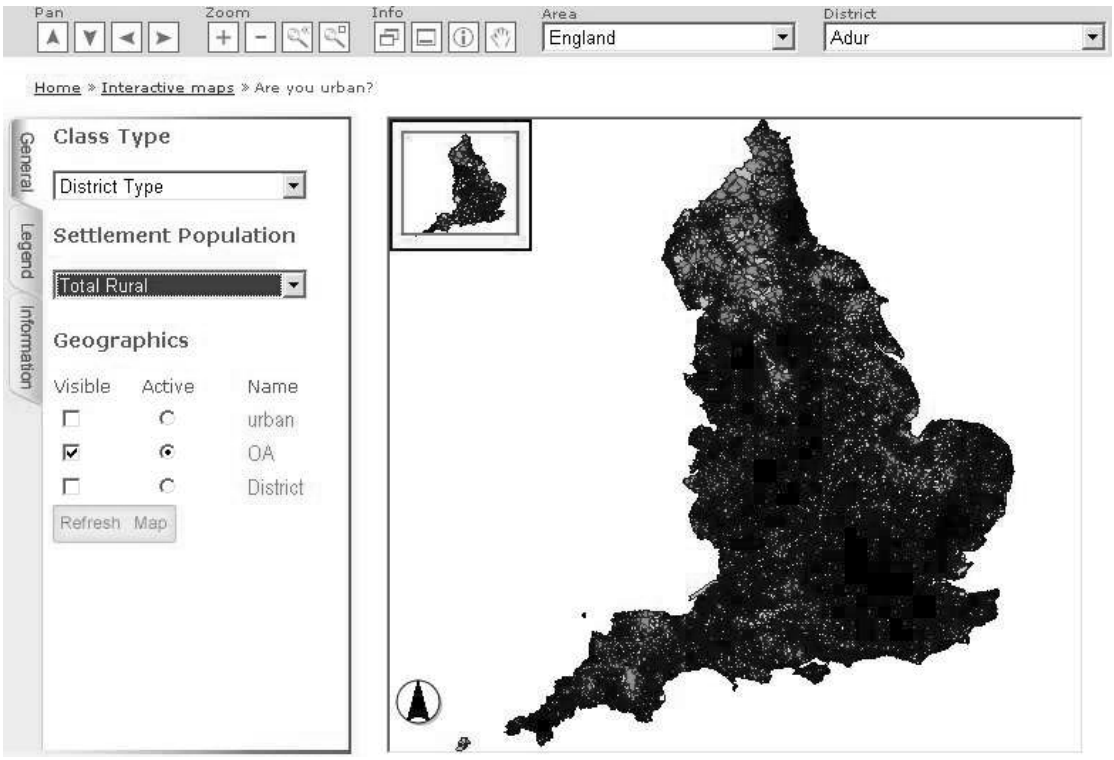



**Figure 2.** Interactive maps section on the RERC website

**Figure 3.** High zoom level of the Rural Urban classification overlaid on Google Satellite map

## 4. Conclusions: the future Rural Urban map mashup

The nature of the Rural Urban definition and the consequent classification of statistical areas entails a series of interesting challenges in visualisation. Building a web map mashup platform seems to be a valuable choice to deploy online the classification (Hudson-Smith et al., 2009 and Haklay et al., 2008). If well designed, a map mashup can be an improvement to the pre existing interactive map available on the RERC website. Firstly because the graphic user interface of typical map mashup is easy to use and incorporates zooming and panning controls that are standard and common among the providers. In

addition map mashups provide general reference data that add context to the data overlaid (see Figure 2 and Figure 4). Contextual data can be aerial photographs (Figure 4) or be displayed using a style similar to the paper version of general reference maps. In particular cases (OpenStreetMap with Mapnik) the cartographic style of the latter type of these contextual data can be changed in order to avoid confusion with the colour schemas of the mashed data. Map mashups often offer alternative ways of visualise attribute data through charts and tables.

A successful map mashup maximises user's experience, but different users have different needs. For the testing purposes the map mashup is designed for general public without a specific knowledge of the Rural Urban Classification. This type of user requires additional care when designing the website layout especially in choosing the initial zoom level at which the data are presented. A proper initial zoom level can help the general public in detect the issues inherent in the way the Rural Urban Classification presented in section 3 -issues that are common when presenting aggregated geographic data (ecological fallacy and modifiable areal unit problem). General public involvement though a user feedback form can help in improving the classification (Singleton and Longley, 2008).

At the moment, the project is still at the initial stage, and different map mashups technologies are being investigated. Some mashups tests of the Rural Urban classification have been created using GMapCreator (Hudson-Smith et al., 2009) and MapTiler (http://www.maptiler.org). GMapCreator developed by the Centre for Advanced Spatial Analysis, University College, London, allows to create a map mashups from different file formats including ESRI Shapefiles. Figure 4 shows one of the map mashup tests, output from GMap Creator.

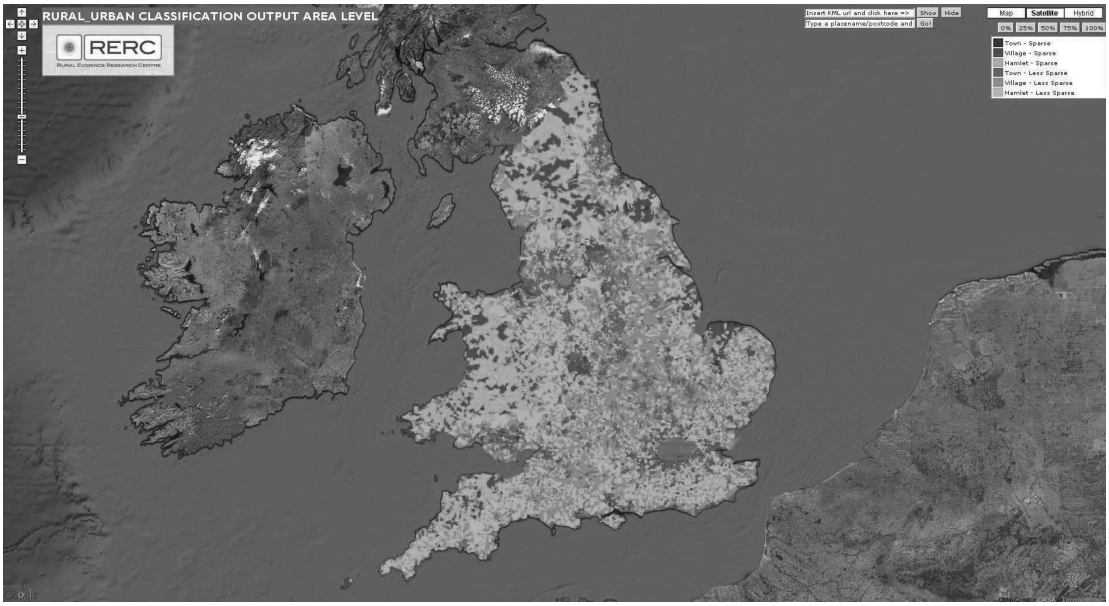



**Figure 4.** GMap Creator's map mashup

Maptiler, is a graphical application for online map publishing based on the GDAL2Tiles libray and developed by Klokan Petr Přidal. Figure 5 shows a mashup of the Rural Urban Classification at Output Area level with the choice of overlaying numerous online base map providers.

Future work will create different mashups for the different aggregation levels although different user surveys confirmed a tendency to use the mashups at very high zoom levels like shown in Figure 4. Additional efforts will be directed to format a page layout that facilitates user experience and embed facilities to add external data, i.e. a Keyhole Markup Language (KML) overlay. New features will also be added to visualise the attribute data through charts and graphs and to create basic summary statistics.

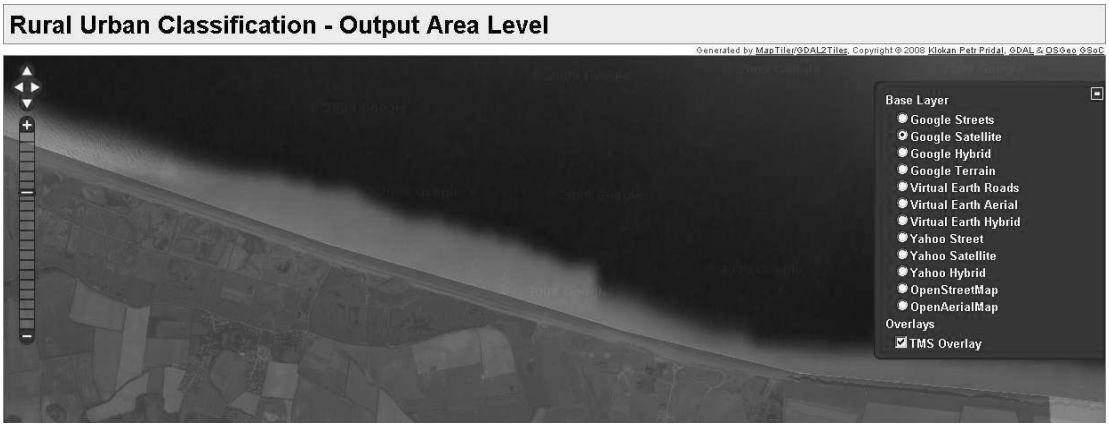



**Figure 5.** Mashup created using MapTiler

## 5. References

Bibby P., Shepherd J. (2004) Developing A New Classification Of Urban And Rural Areas For Policy Purposes -The Methodology, Report To The Office For National Statistics.

Haklay, M., Singleton, A.D., Parker, C. (2008) Web Mapping 2.0: The Neogeography Of The Geospatial Internet. *Geography Compass*. 2(6), 2011–2039

Haklay, M., Zafiri, A., (2008), "Usability Engineering For Gis: Learning From A Screenshot", *The Cartographic Journal*, 45(2) 87-97.

Hudson-Smith A., Crooks A., Gibin M., Milton R., Batty M.(2009) "Neogeography And Web 2.0: Concepts, Tools And Applications", Journal Of Location Based Services, Volume 3, Issue 2, Pages 118 -145, June 2009.

Gibin M., Mateos P., Atkinson P., Petersen J. (2009) Google Maps Mash-Ups For Local Public Health Service Planning In *Planning Support Systems: Best Practices And New Methods*, Stillwell J., Geertman S. Eds, Springer, 2009.

Gibin M., Singleton, A.D., Milton, R., Mateos, P., Longley, P.A.(2008) “Exploratory Cartographic Visualisation Of London Using The Google Maps Api” In Applied Spatial Analysis And Policy, Vol.1 Number 2, July 2008.

SERRL (2002) Review Of Urban And Rural Area Definitions, Report To Office. For National Statistics By Serrl, Birbeck College, Curds, University Of Newcastle, Department Of Town & Regional Planning, Sheffield University.

Singleton A. D., Longley P.A., "Modifying a Geodemographic Classification of the e-Society using public feedback". CASA Working Papers n. 136, June 2008.

Suffolk County Council. (2009) Information Note On The Rural Urban Classification 2004, Available At Http://Www.Suffolk.Gov.Uk (Last Accessed 27/11/2009)

Turner A.J. (2006) *Introduction To Neogeography*, O'Reilly Media.

**Websites** (last accessed 26/11/2009)

http://www.rerc.ac.uk/

http://www.casa.ucl.ac.uk/software/gmapcreator.asp

http://www.klokan.cz/projects/gdal2tiles/

http://www.maptiler.org http://www.suffolk.gov.uk/

## 6. Authors Biography

Maurizio Gibin is Lecturer in GIScience at Birkbeck College, University of London. His research interests are spatial analysis, health geography, geovisualisation ,web cartography and analytical design in

thematic mapping.

John Shepherd is Chair of Human Geography at Birkbeck College, University of London. His main research interests lie in the application of geographical concepts and methods to supporting decision making in public policy, especially in relation to settlement and land use planning.

# Enhancing Environmental Awareness Using Geospatial Mobile Technologies

Hanif Rahemtulla[1]
[1]Department of Geography, University College London, London, UK WC1E 6BT
h.rahemtulla@ucl.ac.uk

KEYWORDS: Environment, Geospatial Web, Mobile Technology, Participation

## 1. Introduction

Since the inception of the environmental movement of the mid 1960s there is growing acknowledgment that 'the global environment is no longer a matter solely for the heads of state and government' (Agenda 21, UN 1992). People's habitats and lifestyle which cannot be changed through the dictate of national strategies, plans, policies and procedures has an impact on local and global environmental problems (Agenda 21, UN 1992). Therefore, the public have a critical role to play in initiating change by contributing ideas and spreading knowledge and involvement (Church and Elster, 2005). 'Indeed, without [public] participation, it is difficult to see how the objectives of Agenda 21 could be reached at all' (UN 2008 p.1). Recently, Prof. Jacqueline McGlade (2008), the Head of the European Environment Agency, called on the need to support the public as well as develop mechanisms that will allow them to play a more active role in the pursuit towards sustainability, especially through the production and consumption of environmental information.

However, expectations from the public about modes of engagement mean that traditional methods of public participation are being challenged. eGovernment systems, such as authoritative Web mapping sites, which were heralded as the solution for over a decade since the emergence of the Web, predominantly offer one-way communication from government bodies to the public and do not include effective means to collect citizen feedback nor engage citizens in two-way dialogue (Rahemtulla and Sieber 2009). The challenge therefore 'is to try to communicate information to people and organizations having different specific concerns, as well as encouraging them to gather and exchange knowledge, and hence participate more in the environmental debate' (Sieber 2007, p.1).

Recent changes in the use of information technology and Web-based resources have provided new opportunities for information dissemination, and, more importantly, for information exchange. Specifically, the emergence of new mechanisms such as the Geospatial Web (GeoWeb) has the potential to address current challenges and build upon current PPGIS/PGIS practice facilitating two-way dialogue between government officials and the public. This opens up new possibilities for communicating and engaging the public on a range of issues, from local environmental inequalities to community action to deal with climate change. For example, mapping platforms such as Google Maps, Platial and Bing Maps allow users to view, share, and contribute user-generated content and volunteered geographic information in an interactive and informative way (Goodchild 2007).

However, while the Geospatial Web provides a rich tapestry of information considerable skepticism remains regarding its use as a mechanism to enhance meaningful communication amongst stakeholders (Keen 2007). This arises, at least in part, as the Geospatial Web reinforces and extends existing barriers to ICT engagement. As Ellul *et al.* (2009a) notes, to access the full functionality of digital map platforms, Internet users require a high bandwidth connection and an

implied level of spatial and digital literacy. From a UK perspective Changing Media (2007) estimate that 2.7 million households have narrowband access to the Internet. Further, Skarlatidou and Haklay (2006) and Nivala (2008) have shown that the success rate in operating and navigating even simple public mapping sites is limited to between 60 and 80 percent.

Mobile phones are an example of a New Information and Communication Technology (NICT) which may overcome issues to digital engagement while extending access to information currently held in Geospatial Web-based applications (see Ellul *et al*. 2009b). The UK has one of the highest levels of mobile phone ownership in Europe (OfCom 2007). Further, data-driven mobile phone services are seen as part of a wider toolkit by which to engage, interact and empower communities (Katz and Aspden 1998). As Olsen (1999) notes, such services will allow individuals and communities to interact with people and information in a myriad of ways.

This paper presents the background to, and description of, a geospatial mobile service (EcoTEXT) designed for an environmental organization (London 21) to further the dissemination of information held in community-driven geospatial web-based services and increase environmental awareness and community engagement.

**2. London 21**

London 21 is an umbrella environmental organisation that works across the public, private and voluntary sectors between different communities and all faiths with the objective of encouraging London's grassroots and community-based organisations that are working towards sustainability to promote themselves and network London-wide for a greener, cleaner and more equitable London.

Today, London 21 has links to over 1500 active grassroots and community-based organizations (London 21 2008). The scale of the network is a testament to the accomplishments of London 21, both in terms of its contribution to capacity building and networking. For instance, London 21 has developed co-ordination mechanisms that support cohesive action and provides a range of services, including training and networking events tailored to local demands and conditions. One example is the London Sustainability Weeks (LSW) (Figure 1).


**Figure 1.** The LSW are held during the first two weeks of June to coincide with World Environment Week. It provides over half a million of London's residents, workers and visitors with the opportunity to discover and experience the diversity and creativity of the

hundreds of initiatives, projects and organisations occurring throughout London that are making a real contribution to a more sustainable capital (London 21 2007).

As is typical of many such organisations, London 21 is built on its diversity and channels of communication. The provision of information through NICTs is at the heart of the organisations strategy to promote, network and support grassroots and community-based organisations in London. To date, London 21 has commissioned the development of several community-driven geospatial web-based services including the London Green Map, Love London and Mapping Change for Sustainable Communities (see Table 1).

**Table 1.** London 21's Geospatial Web Services.

| **London 21's Geospatial Services** | **Description** |
|---|---|
| London Green Map | An online interactive map for Londoners that highlights over 1200 community projects and services for sustainable living. The aim is to make available in visual, attractive form local information to help people live in a greener, healthier way, with a particular emphasis on the creativity of community action. The types of projects included on the map includes: farmers' markets, children's playgrounds, community groups and event, local charities, skills and training opportunities, theatre workshops, and recycling centres. |
| Love London | This is the official LSW website showcasing projects from major campaigns to locally-based initiatives occurring as part of the festival. The site acts a central hub where event organisers (e.g. grassroots and community-based organisations) can register and display their event on the site and where individuals can explore, contribute and access information about those events occurring across London |
| Community Maps | An online interactive GIS-based map of East London, the Lower Lea Valley and the Thames Gateway (Figure 5.7). The Map site is based on the existing London Green Map, and provides two main interfaces for users– an overview map of the East London and the Thames Gateway and a series of community maps within this region providing users with access to detailed information about their local area. The Map also provides a virtual environment in which communities can record new significant developments in their areas as well as highlight development sites, environmental issues and projects, local issues and stakeholder groups. |

In line with the objectives of the network, the organisation is continually seeking new ways of extending access to London 21's extensive information about green projects and services to a new audience and overcome existing barriers to NICT engagement. This led to the development of a mobile service EcoTEXT. This service allows London 21 to automatically send text-messages to registered members notifying them of events occurring in their local area. Further, the service can be accessed using any kind of mobile terminal, regardless of the network, operator, or platform to which the terminal is connected.

**3. EcoTEXT**

EcoTEXT is an example of a locally based and locally driven information and communication system. This service allows individuals to receive geographically targeted, action-orientated, time-relevant information via text messages on their mobile phones. The content of the service is information about upcoming local environmental events and activities, which match the interests to the user, when these events occur in close spatial proximity to where that user resides (Figure 2). This type of service represents a powerful new dimension for the provision of data-driven services in comparison to current text-based services; relating location to information and giving the service additional meaning and value (see Rahemtulla *et al*., 2008).


**Figure 2.** EcoTEXT - A Mobile Spatial Messaging Service

**4. Community Engagement and Participation**

Exploratory studies of the service were conducted during LSW 2007 and as part of the MCSC project 2008. For each study, individuals could subscribe to receive free targeted messages about environmental initiatives occurring across the capital. Individuals could subscribe to the service either completing an online registration form or a paper-based version of the form. The use of the paper-based form was significant in this study to allow non-Internet users to register to the service. The service was promoted to the widest possible audience through the press and mass media.

The EcoTEXT service was subscribed to by 43 individuals. Of these, 38 individuals received one or more messages with a total of 198 text-messages being sent. While the level of adoption to this service would at first glance seem low, it can be explained by the technical difficulties experienced and subsequent delays in the implementation of the service.

For each exploratory study, the service was subscribed to by individuals characterised as ‘early adopters’ of technology. This can be explained by the technology adoption and diffusion patterns within communities and the strategy of introduction which was followed. While these early adopters are by no means representative of the general population they are by far the most important segment since the mass market copies their behaviour and product usage (see DeMarez and Verleye 2001). Therefore their perceptions towards and experiences of this service are relevant as they have an important ‘opinion leader-role’ to play towards the rest of the market. The exploratory studies reveal that early adopters to the service had a positive experience of EcoTEXT in terms of its usability, practically and usefulness. More importantly, early adopters

found that the service fulfilled a real purpose and function which as Rogers (1995) notes, is essential; for if there is no reason to use the technology it will be quickly discarded as irrelevant. The positive experiences of early adopters will lay the foundation upon which to bring the mass market on board – crossing the chasm between the early adopters and early majority (see Rogers 1995).

Further, the introduction of this service into the organisation's communication toolkit has shown to create, supplement and strengthen social ties and interactions within the community. The study found that the service EcoTEXT increased levels of civic engagement with 87 percent of subscribers forming new contacts and connections with the community; 85 percent stating an increase awareness of local environmental events and 77 percent participating in an event or activity based in information they received through the service.

## 5. Conclusions

EcoTEXT complements and extends access to London 21's existing NICT web-based services and information. Further, from London 21's perspective, where the emphasis is upon engaging with the widest possible audience, the service can be accessed using any kind of mobile terminal, regardless of the network, operator, or platform to which the terminal is connected.

While the underlying technology for this service, SMS, has been tried and tested in the past and as such, is not cutting edge on its own, the service is unique in its application. EcoTEXT is a highly personalised mobile service leveraging locational information. Further, in contrast to existing mobile services, EcoTEXT disseminates genuine local content, encouraging user input, as opposed to end users passively receiving information.

Further, the introduction of such a service into London 21's range of communication tools has shown to supplement and strengthen local social network ties, information exchange, and encourage civic engagement ranging from informal social interaction through to awareness of events leading to, it is anticipated, an increase in face-to-face interactions. Furthermore, as the reach of the service expands, the possibility exists by which to engage with individuals which where previously unconnected thus, strengthening and even initiating weak ties within the community (see Haythornwaite 2005). The cumulative affects of this service will it is hypothesized, be to assist in bringing collocated people together socially and increase awareness of individual community skills and assets.

## 6. References

Changing Media. (2007). *UK Internet Usage – A Report for the BBC Trust*. Available from:http://www.bbc.co.uk/bbctrust/assests/files/pdf/regulatory_frameworkservice/research.pdf [Accessed 1 January 2008].

DeMarez, L. and Verleye, G. (2001). *Diffusions of Innovations: Successful Adoptions needs more effective targeting*. Available from: http://www.escp-eap.net/conferences/marketing/pdf_2002/demarez.pdf [Accessed 1 January 2008].

Ellul, C., Rahemtulla, H.A, and Haklay, M. (2009). *Beyond the Internet – Increasing Participation in Community Events by Text Messaging*, Paper presented at the 27th Urban Data Management Symposium (UDMS), Ljubljana, Slovenia, June 24-26.

Ellul, C., Haklay, M., Francis, L. and Rahemtulla, H.A. (2009). *A Mechanism to Create Community Maps for Non-Technical*, Paper presented at the International Conference on

Advanced Geographic Information Systems & Web Services (GEOWS), Cancun, Mexico, February, 13-15.

Goodchild, M.F. 2007. “Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0” *International Journal of Spatial Data Infrastructures Research* 2: 24–32.

Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report Working Group II 2007. Impacts, Adaptation and Vulnerability. Cambridge University Press.

Keen, Andrew. 2007. *The Cult of the Amateur: How today's internet is killing our culture.* Doubleday, New York, NY, USA.

London 21. (2008). *London 21 Sustainability Network* [online]. London 21. Available from: *http://www.london21.org* [Accessed 02 January 2008].

Moore, G.A. (1991). *Crossing the chasm: Marketing and selling technology products to mainstream customers*. New York: Harper Business.

OfCom. (2006). *The Communications Market 2006: Nations and Regions Research Report*. London: Office of Communications.

Rahemtulla, H., Haklay, M. and Longley, P.A. (2008). *A Mobile Spatial Messaging Service for a Grassroots Environmental Organisation*. Journal of Location Based Services 3(1) 28-52.

Rogers, E. (1995). Diffusion of Innovations. New York: The Free Press.

Sieber, R. and Rahemtulla, H.A. (2009). *The Participatory Geoweb,* Paper presented at Royal Geographic Society (RGS-IBG) Annual Conference, Manchester, August 22-23.

Skarlatidou, A. and Haklay, M. (2006). *Public Web Mapping: Preliminary Usability Evaluation*. Paper presented at GISRUK 2005, Nottingham, April 5-7.

Nivala, A.M., Brewster, S.A. and Sarjakoski, L.T. (2008). *Usability Evaluation of Web Mapping Sites*. The Cartographic Journal, Use and Users Special Issue, 45(2) 129-138.

**Acknowledgements**
This research was sponsored by the Economic Social Research Council (ESRC) and HEIF UrbanBuzz Programme with further financial support from University College London.

**Author**
Dr. Hanif Rahemtulla is an Honorary Research Fellow at University College London. His focuses on the use of computer mediated communications to engage the public on the crucial issues of our time, such as global environmental change and sustainable development.

# Where's the analysis? Evaluating the OGC's Web Processing Service

Nick Gould

Greater Manchester Transportation Unit, Manchester City Council, UK
Tel. +44 161 455 2140 | Email: n.gould@manchester.gov.uk

KEYWORDS: Web Processing Service, OGC, remote geoprocessing, GeoWeb, road accidents

## 1. Introduction

The development of the geoweb has facilitated the development of large variety of mapping applications where public and private datasets are overlaid on fast, slippy base maps providing a rich visualisation experience. Few of these web applications, however, provide any spatial analysis tools.

One possible solution is the Open Geospatial Consortium's Web Processing Service (WPS), an open standard that defines a protocol for remote geoprocessing using a web service (Open Geospatial Consortium, 2007).

The aim of this paper is to investigate the potential of the WPS by using road traffic accident data as a case study. The intention is to develop a WPS service that will highlight accident black spots by identifying clusters of accidents.

## 2. The Web Processing Service

The WPS is part of the OGC's Open Web Services (OWS) suite of protocols for geospatial web services. OWS includes the Web Mapping Service (WMS) for delivering map images and the Web Feature Service (WFS) for delivering vector data. A number of commercial and open source GIS applications are now WMS and WFS compliant (Open Geospatial Consortium, 2009).

The WPS can be seen as a logical progression from WMS and WFS. A WPS implementation can perform any number of spatial analysis tasks. Source data is supplied to the WPS service, processed, and the results returned to the client. The protocol defines how data is to be supplied and how the results are returned but does not mandate how the processing is done. Output may be anything from a single number to a GML file to a JPEG.

The WPS has been identified by Foerster and Stoter (2006) as a way of exchanging knowledge by sharing and reusing algorithms. As new algorithms are developed they can be made easily available to other researchers in the field. In addition the WPS offers platform independence and the capacity for integration with existing applications (de Jesus, et al., 2008). Friis-Christensen, et al. (2007) identify decentralization as another benefit of the WPS.

However, possible disadvantages of the WPS have been identified. The protocol does not specify a notification mechanism for informing a client when the processing is complete. Foerster and Stoter (2006) regard this as a particular problem for a long-running process, where the client has to repeatedly check the status of the process. De Jesus, et al. (2008) identify the use of GML as input to a WPS service as potential bottleneck given the verbosity of XML and GML in particular.

WPS applications have been developed for assessing forest fire damage in Europe (Friis-Christensen, et al., 2007); predicting water runoff (Díaz, et al., 2007); generalising road networks (Foerster and Stoter, 2006); and spatial interpolation (de Jesus, et al., 2008).

## 3. Methodology

To assess the WPS protocol, a demonstration WPS was developed. A web client for the service was then built.

The dataset to be analysed consisted of road traffic accidents in Greater Manchester. The dataset holds approximately 108,000 records of point data. There are a number of different techniques for identifying black spots of road accidents (Geurts and Wets, 2003). The method chosen was to use a density-based algorithm, DBSCAN as developer by Ester, et al. (1996). The particular implementation of DBSCAN employed was that described by Tan, et al. (2006) who optimised the original algorithm for simplicity. The algorithm attempts to identify clusters of points in a given dataset; points are either added to a cluster or discarded as noise. DBSCAN has two parameters *Eps*, which is the distance at which two points are close enough to be in the same cluster, and *MinPts*, which is the minimum number of points in a cluster. Three types of points are defined: *core* points, where the number of points within a distance, Eps, of the point is greater than or equal to MinPts; *border* points which fall within Eps of a core point; and noise points, the remainder. Ester, et al. (1996) concluded that MinPts could safely be set to 4 for all 2D datasets; a decision that Tan, et al. found "reasonable" (p529).

## 4. Implementation

The source data was imported into a Windows Server installation of *postGIS*. The *GeoServer* application was installed on the same server to deliver the data source as a Web Feature Service (WFS). The WFS acts as an input to the WPS; specifically a WFS request will be included as one of the input parameters to the WPS which will be passed to the DBSCAN algorithm. To demonstrate the distributed nature of OWS, the WFS was implemented on a separate server to the WPS (Figure 1).

The WPS was developed using the model suggested by Michaelis and Ames (2007) in their early evaluation of the WPS. They suggest building a wrapper, a piece of code that will act as an interface to the geoprocessing routines (in this case the DBSCAN algorithm). The DBSCAN algorithm was implemented by an ASP.NET web service. The WPS interface, which handles all requests to the service (Figure 1, Figure 3) was implemented using a standard ASP.NET web application.

Geoprocessing operations involving large datasets can be time consuming, taking several minutes or more to complete. Since a request to a web server times out after a few minutes, a client cannot be expected to wait for the results. Any implementation needs to be able to operate in an *asynchronous* manner. That is, an *Execute* request has to be made and a response acknowledging the request returned immediately without waiting for the results. A request is made for the results in a separate step.

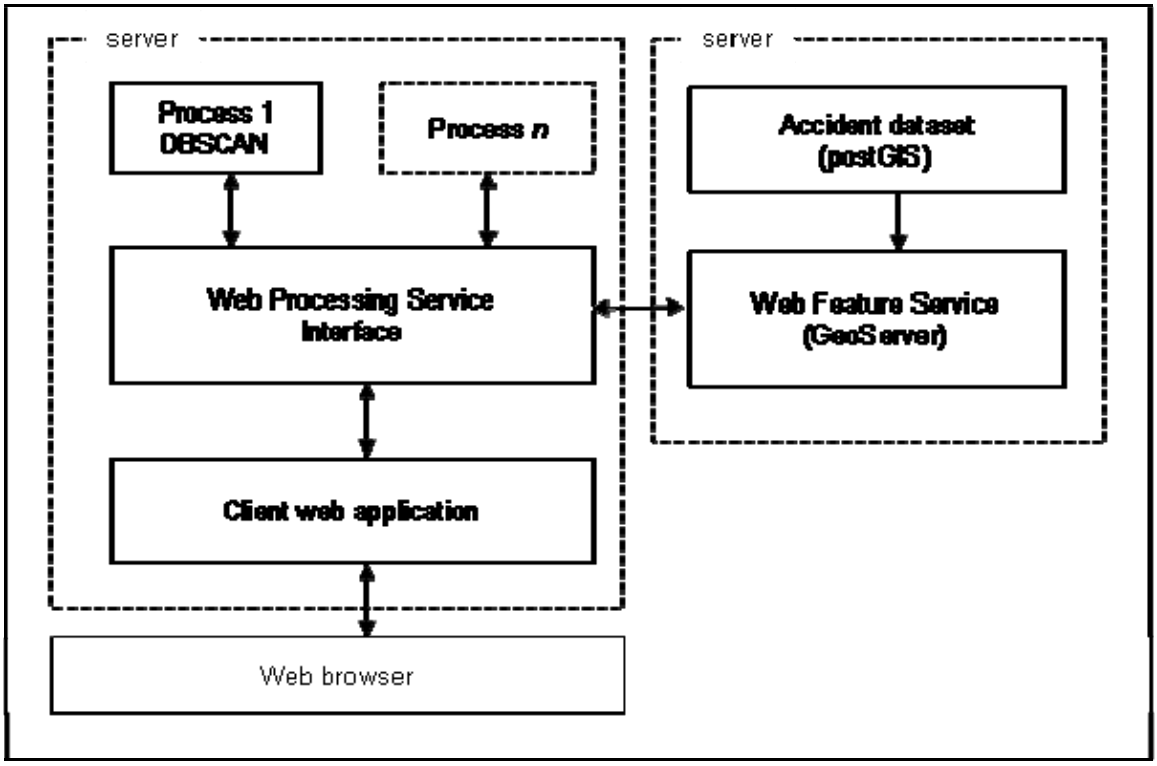



**Figure 1** Implementation

Any request to the WPS returns a *response document*, an XML file stored on the server (Figure 2). If the request fails for any reason then the response document will contain an error message. If a successful *Execute* request is made then the response document will contain a reference to the results file. The results are written to a Keyhole Markup Language (KML) file on the server. The advantage of writing the results to a file is that the results can be retrieved at a later date without re-running the request.

```
- <wps>
  - <ProcessSucceeded>
    - <ExceptionText>
        Success! Total points 398 with Eps = 20 163 core points 37 border points 198 noise points Points in a cluster: 200 number of clusters =19
      </ExceptionText>
      <processID resultsFile="http://www.gmtu.gov.uk/wps/output/DBSCAN2009618135243778.kml">2009618135243778</processID>
    </ProcessSucceeded>
  </wps>
```

**Figure 2** The response document

Since the service operates in an asynchronous manner, and there is no mechanism in the WPS protocol to inform the client that the results are ready, the client has to poll the response document until the results are ready. The time taken to generate the results file is a function of the number of points being analysed and the loading on the server(s).

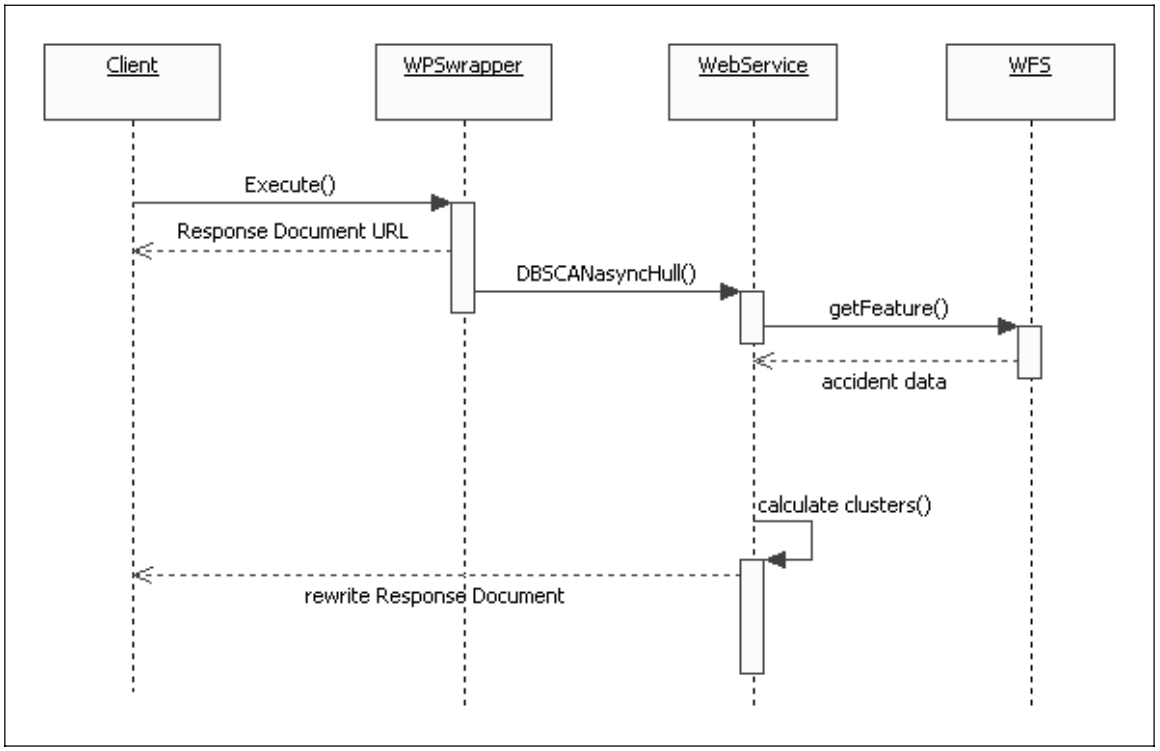



**Figure 3** UML sequence diagram for an Execute request

## 5. Developing the client

An *Execute* request to the WPS can be made by entering a URL into a web browser. However, the URL is long and complex; in particular the section that specifies the call to the WFS requesting input data. The WPS can be better accessed by developing a web client application.

The application developed allows the user to draw a bounding box on a Google Maps window (Figure 4) and then display the KML file of accident clusters in the specified area. The clusters will be displayed as points (Figure 5) or as polygons (Figure 6) depending on an input parameter supplied to the WPS.

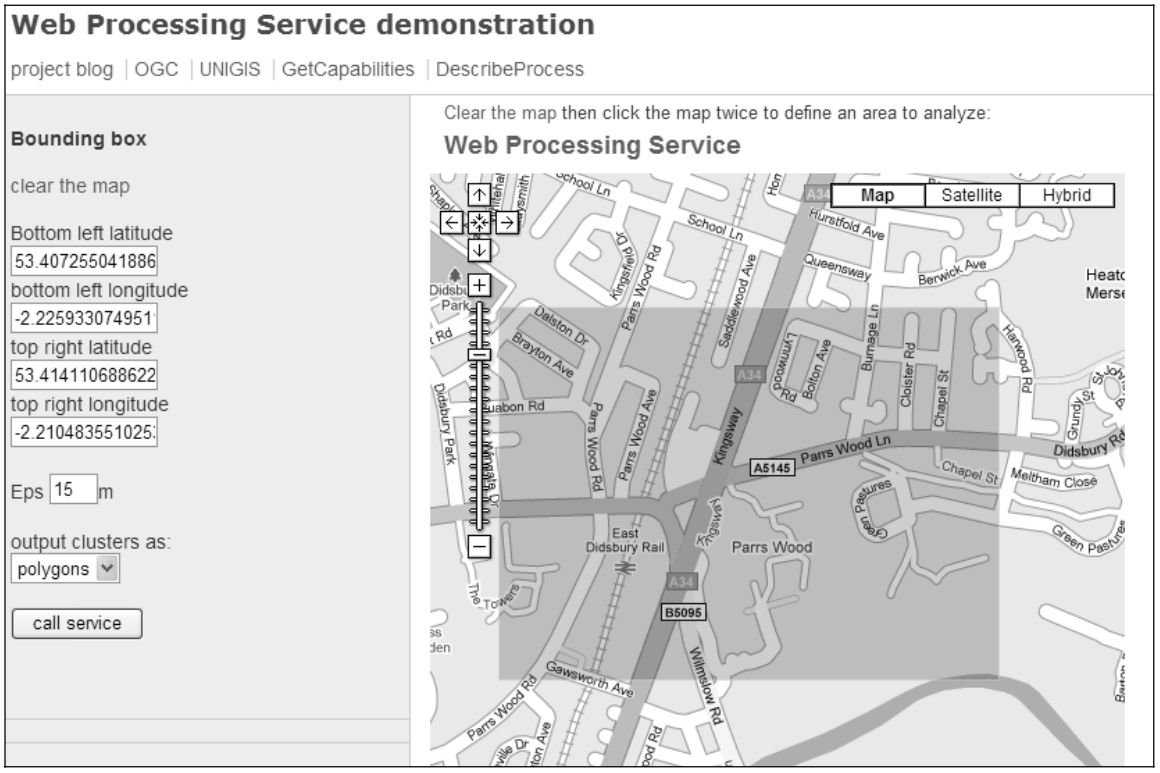



**Figure 4** WPS client - defining the bounding box

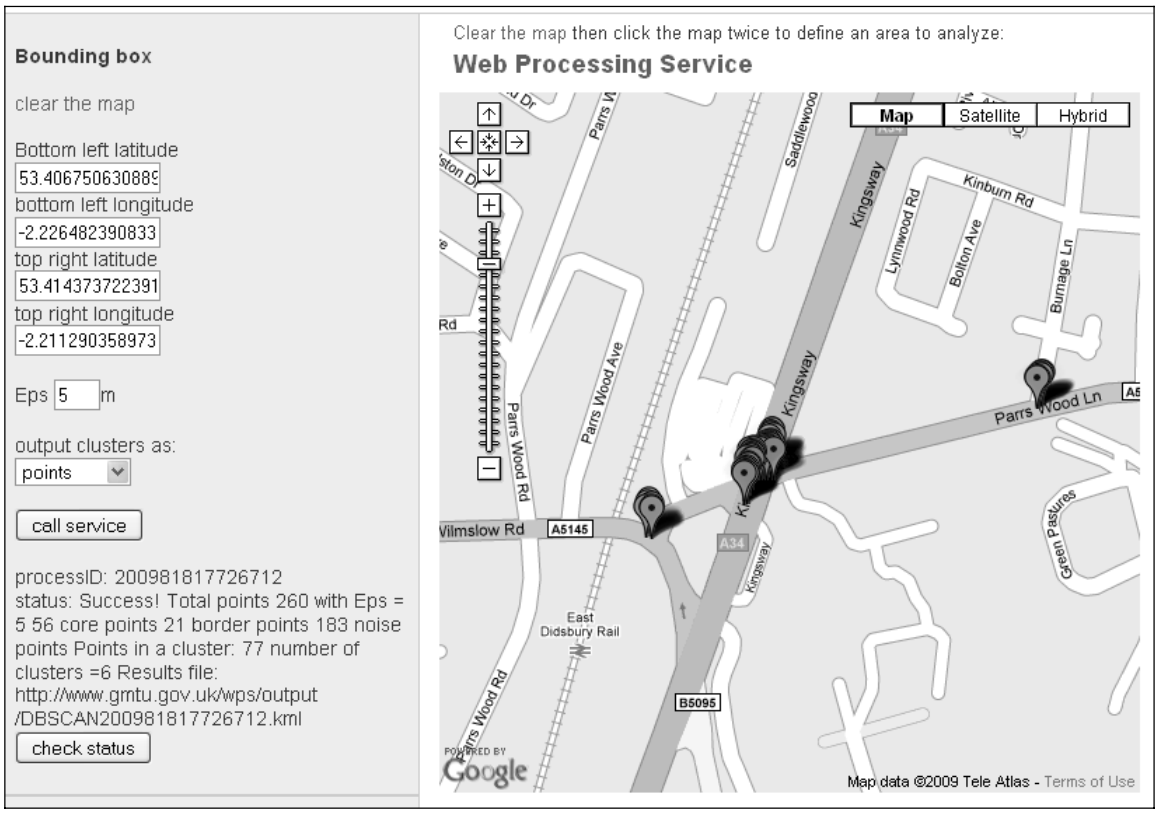



**Figure 5** WPS client - displaying the results as points


**Figure 6** Accident clusters as polygons

## 6. Conclusions

This paper has demonstrated the capability of the WPS to provide remote processing of spatial data. The WPS is unlikely to replace the traditional desktop GIS but, as suggested by Kiehle et al (2006), it may have a role in Spatial Data Infrastructures (SDI) where large datasets are stored remotely. The WPS removes the need to download large datasets for local processing.

## 7. Acknowledgements

I am grateful to the Greater Manchester Transportation Unit for permission to use their road traffic accident dataset. The work was undertaken for an MSc dissertation at UNIGIS and I would like to thank my supervisor Derek Reeve of the University of Huddersfield for his guidance.

## References

de Jesus, J., Dubois, G., Hiemstra, P. (2008). Web-based geostatistics using WPS. *GI Days 2008*. University of Münster, Germany, 16-17 June 2008.

Díaz, L., Costa, S., Granell, C., Gould, M. (2007) Migrating geoprocessing routines to web services for water resource management applications. *10th AGILE International Conference on Geographic Information Science*. Aalborg University, Denmark, 8-11 May 2007. Available at: http://www.plan.aau.dk/~enc/AGILE2007/PDF/151_PDF.pdf [Accessed 24 November 2009]

Ester, M., Kriegel. H., Sander J., Xu, X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *2nd International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon, USA, August 1996. AAAI Press.

Foerster, T., Stoter, J. (2006) Establishing an OGC Web Processing Service for generalization processes. *Workshop of the ICA Commission on Map Generalisation and Multiple Representation*. Portland, USA, 25th June 2006.

Friis-Christensen, A., Ostländer, N., Lutz, M., Bernard, L. (2007) Designing Service Architectures for Distributed Geoprocessing: Challenges and Future Directions. *Transactions in GIS*, 11(6), 799-818.

Geurts, K., Wets, G. (2003) *Black Spot Analysis Methods: Literature Review*. Policy Research Centre Mobility & Public Works, Hasselt University, Belgium. [Online] Available at: www.steunpuntmowverkeersveiligheid.be/nl/modules/publications/store/13.pdf [Accessed 22 November 2009]

Kiehle, C., Greve, K., Heier, C. (2006) Standardized Geoprocessing - Taking Spatial Data Infrastructures one Step Further. 9th AGILE Conference on Geographic Information Science. Visegrád, Hungary, 20-22 April 2006.

Michaelis, C. D., Ames, D. (2008) Evaluation and Implementation of the OGC Web Processing Service for Use in Client-Side GIS. *Geoinformatica*, 13(1).

Open Geospatial Consortium (2007) *OpenGIS Web Processing Service*. [Online] Available at: www.opengeospatial.org/standards/wps [Accessed 24 November 2009]

Open Geospatial Consortium (2009) *Implementing Products*. [Online] Available at: www.opengeospatial.org/resource/products/implementing [Accessed 24 November 2009]

Tan, P., Steinbach, M., Kumar, V. (2006) *Introduction to Data Mining*, Addison-Wesley.

**Biography**

Nick Gould works for the Greater Manchester Transportation Unit, a local government unit based at Manchester City Council. He has recently completed an MSc in Geographic Information Science at Manchester Metropolitan University (UNIGIS).

# Describing Spatial Relations using Informal Semantics

Kristin Stock[1]

[1]University of Nottingham, NG7 2RD, Nottingham
Tel. +44(0)787 299 3270
Email: kristin.stock@nottingham.ac.uk
Web address: http://www.nottingham.ac.uk/~lgzwww/contacts/staffPages/kristinstock/index.html

KEYWORDS: spatial relations; semantics; linguistics; natural language

## 1. Introduction

Significant work has been undertaken towards the development of formal, mathematical and precise definitions of spatial relations (Egenhofer et al 1994; Cohn et al 1998). However, most people do not conceptualize spatial relations in such a way. In contrast, people tend to think in terms of **informal semantics**. Informal semantics take the form of cognitive models that are experiential or naïve rather than formal and precise.

Although there are likely to be similarities in the cognitive meanings of spatial relations between individuals due to common perceptual and linguistic experiences (Mark et al 1996), such models are also likely to be individually distinct, culturally-enmeshed, vague, dynamic and context sensitive. Formal approaches to the definition of spatial relations do not allow these kinds of variation to be reflected, and also are not convenient for immediate use by non-experts. Such approaches usually require specialist knowledge and understanding of a formal definitional language, and do not coincide well with the informal semantics used in an everyday context.

Various attempts have been made to resolve this mismatch between the formal approach to the definition of spatial relations and the more informal ways in which humans think. Prominent among these is research on fuzzy geographic concepts and regions (for example, Schockaert et al 2006; Dilo et al 2007); work developing approaches that allow for variations in the parameters that define a geographic concept (Bennett et al 2008); study of the ways in which people interpret vague geographic notions like 'downtown', also known as vernacular geography (Montello et al 2003; Waters et al 2003); examination of cultural and multi-lingual differences in the meaning of geographic features (Mark 1993) and mathematical modelling of notions like 'near' and 'far' using fuzzy approaches (Robinson 1990; Robinson 2000).

All of these previous approaches either try to formally define the semantics of spatial relations, or take a more informal approach that is based around asking users to define the extension of a concept or a relation, rather than the intention (definition). These approaches normally aim to find a commonly held, consensus view of a particular spatial relation or concept, rather than attempting to capture individual variations.

This paper presents the results of an experiment that evaluated the use of Natural Semantic Metalanguage (NSM) for the definition of spatial relations by non-experts, and the consequent comparison of those semantic definitions. This approach has the benefits of allowing informal semantics to be expressed and capturing individual variations in semantics, rather than requiring a consensus view to be adopted.

NSM is a simple language that has developed from a body of linguistics research over the last 35

years and has defined a set of semantic primitives: 63 words or phrases that have been found to exist in all of the many human languages that have been studied by linguists working in the area over the last few decades; that are essential for the explanation of the meaning of other words and grammatical constructions and that cannot themselves be explained in a non-circular fashion. Table 1 shows these primitives (Goddard 1998; Goddard et al 2002).

**Table 1.** NSM Semantic Primitives

| | |
|---|---|
| Substantives: | I, YOU, SOMEONE, PEOPLE, SOMETHING/THING, BODY |
| Relational substantives: | KIND, PART |
| Determiners: | THIS, THE SAME, OTHER |
| Quantifiers: | ONE, TWO, SOME, ALL, MUCH/MANY |
| Evaluators: | GOOD, BAD |
| Descriptors: | BIG, SMALL |
| Mental predicates: | THINK, KNOW, WANT, FEEL, SEE, HEAR |
| Speech: | SAY, WORDS, TRUE |
| Actions, events, movement, contact: | DO, HAPPEN, MOVE, TOUCH |
| Location, existence, possession, specification: | BE (SOMEWHERE),THERE IS / EXIST, HAVE, BE (SOMEONE / SOMETHING) |
| Life and death: | LIVE, DIE |
| Time: | WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT |
| Space: | WHERE/PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE |
| "Logical" concepts: | NOT, MAYBE, CAN, BECAUSE, IF |
| Intensifier: | VERY, MORE |
| Similarity: | LIKE |

The paper describes an experiment and its outcomes, and in doing this, shows how NSM can be used to compare the semantic equivalence of expressions constructed by different individuals, even if they are non-experts using naïve, non-literal, individually-specific definitions of the spatial relations concerned.

## 2. Method

We asked 6 participants to write NSM definitions for 4 common spatial relations. The method adopted for the experiment involved the presentation of a short series of instructions to participants. These instructions briefly explained NSM, providing simple examples and basic rules. Participants were then asked to construct definitions of four spatial relations (*contains, next to, on* and *intersects*) using only the NSM semantic primitives (from Table 1). The experiment was undertaken in English.

The experiment used a qualitative and not a quantitative approach, and as such did not aim to examine a statistical or representative sample of people in order to gain an indication of the full range of different types of definitions that people might make, or to generalise across the entire population of humans within a culture. The aim was to examine in detail a small number of participants, to explore the variations that were evident among this group and to examine the ways in which non-experts constructed definitions of spatial relations using NSM. Given these aims, the small number of participants provided adequate material for useful results, showing a range of different approaches to the construction of NSM definitions. Furthermore, as a qualitative study with such aims, the participants were not intended to form a random sample, although variety in professional background was sought and achieved.

## 3. Results

Table 2 shows the actual NSM definitions that the participants wrote during the experiment (definitions shown in grey are not valid NSM). As can be seen, there is significant variation among the definitions provided, indicating that 6 participants still provide useful material that can be used to derive equivalence rules (see Section 4). This does not claim that with additional participants a greater range of expressions would not be found, and more experiments are likely to be undertaken in the future.

**Table 2.** Experiment Results

| Respondent | *Next to* | *On* | *Contains* | *Intersects* |
|---|---|---|---|---|
| 1 | A NEAR SIDE WHERE B LIVES. | A EXISTS IN THIS PLACE WHERE B EXISTS. | ALL B [IS] PART OF A. | A CAN TOUCH B FOR A SHORT TIME. |
| 2 | A NEAR B. | A ABOVE B. | A INSIDE B. | A KIND OF B. |
| 3 | SOMEONE EXISTS HERE IN THIS PLACE VERY NEAR SOMETHING OTHER. | THERE IS SOMETHING ABOVE SOMETHING BELOW. | INSIDE THIS BODY THERE IS SOMETHING OTHER. | SOMEONE EXISTS IN THIS PLACE, IF NOT NEAR, WHERE THERE IS SOME OTHER; MAYBE FOR SOME TIME, MAYBE A SHORT TIME |
| 4 | IN THIS PLACE THERE IS A NEAR B. | THERE IS A ABOVE B. A TOUCHES B. | B EXISTS INSIDE A. | FOR SOME TIME PART OF A EXISTS NOT LIKE B. NOW A EXISTS IN THIS PLACE WHERE B EXISTS. |
| 5 | I LIVE HERE NOW. YOU LIVE NEAR THIS PLACE. YOU HAPPEN TO LIVE THE SAME SIDE. | I LIVE ABOVE YOU. YOU LIVE BELOW THIS PLACE. | THERE IS SOMETHING SMALL INSIDE THIS BODY. SOMETHING NOT THE SAME. | PART OF THIS THING TOUCHES THIS OTHER THING IN THIS PLACE. |
| 6 | A IS NEAR B. | A IS ABOVE B. | A IS INSIDE B. | A IS BELOW ABOVE B. |

The variations in the definitions provided for each given spatial relation can be divided into two categories:

- variations in expression and
- variations in semantics.

Variations in expression represent different ways in which participants have chosen to describe the spatial relations but do not indicate semantically different concepts. Variations in semantics describe actual semantic differences. A simple example of a variation in expression that is not a variation in semantics can be seen in the comparison of: 'A INSIDE B' and 'A EXISTS INSIDE B'. In contrast, 'THERE IS SOMETHING INSIDE THIS BODY' is semantically different from 'THERE IS SOMETHING SMALL INSIDE THIS BODY', in that the second definition also requires that the thing inside is small.

In order to use this method effectively to examine the differences in the informal semantics held by individuals for the selected spatial relations, the differences in expression must be isolated from the differences in semantics. The latter are important while the former can be viewed as 'noise' that distracts from the comparison of the conceptual models held by individuals. To this end, a number of equivalence rules were defined to express differences in expression that were not also differences in semantics and could thus be removed for the purposes of semantic comparison.

## 4. Equivalence Rules

The definitions provided by participants were examined in detail in order to identify differences in the ways that participants constructed their definitions that did not necessarily reflect semantic differences. On this basis, the following equivalence rules (1) to (7) were defined. These rules are of particular relevance in the spatial context, either because they directly address spatial information, or because they relate to items that are often used to describe spatial relations.

$$\text{IS (SOMEWHERE)} \equiv \text{THERE IS/EXIST} \equiv \text{LIVES} \quad (1)$$

$$\text{WHERE} \equiv \text{IN THIS PLACE} \equiv \text{HERE} \equiv \text{`\ `} \quad (2)$$

$$\text{BODY} \equiv \text{A} \equiv \text{B} \equiv \text{SOMETHING/THING} \equiv \text{SOMEONE} \equiv \text{I} \equiv \text{YOU} \quad (3)$$

$$\text{TEMPORAL PRIMITIVES \{WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT\}} \equiv \text{`\ `} \text{ if mentioned only once in the definition} \quad (4)$$

$$\text{SPATIAL PREDICATE} \equiv \text{IS} + \text{SPATIAL PREDICATE} \quad (5)$$

$$\text{THIS} \equiv \text{`\ `} \quad (6)$$

## 5. Determining Equivalence

The four spatial relations were compared using the rules above remove differences in expression ('noise') and allow more detailed comparison of the semantic variations between individuals. This Section describes the process for the *On* spatial relation as an example.

The semantics of the *On* spatial relation have been discussed at length by NSM linguistic experts (Goddard 2002), including different semantic possibilities like 'the key is on the chain' and 'fins on its back'. However, participants were given no more guidance than that the experiment was about language and space, and most described the typical spatial relation. Five participants provided valid definitions, and their equivalence can be evaluated as follows:

O6 ≡ O2 (Rule 5)
O3 ≡ O5 (Rule 3, Rule 1)
O6 ≡ O3 (Rule 3, Rule 1), additional semantics: BELOW
O6 ≡ O4 (Rule 1), additional semantics: TOUCHES

Three of the definitions provide additional semantics. Two of them draw the distinction between ABOVE and BELOW by including both primitives. A third explication includes TOUCHES to describe the *On* relation. These more specific descriptions indicate different levels of specificity in the participants' thinking or expression and are semantically interesting.

## 6. Discussion and Future Work

A number of observations arose from this work:

1. It is clear from the experiment that simple words for spatial relations are not understood in a common way by non-experts. The *intersects* relation is an example of this.
2. There are numerous examples in the responses of what has sometimes been called naïve geography. The described notions are often vague (for example, few participants described *next to* as anything more than nearness, even though most people would intuitively consider it to include the idea of being beside (which is available as a semantic primitive).
3. In addition to vagueness in the notions themselves, participants varied in their style of expression.

This paper has identified that some differences in the ways in which NSM definitions are written represent real differences in semantics, while others are only differences in expression. It is the assertion of this paper that the limited grammar and vocabulary of NSM dictates that it is possible to define a set of rules that would allow the semantic equivalence of NSM definitions to be evaluated and thus allow differences of expression to be identified and removed, clearing the way for a more thorough examination of differences in semantics. This paper has illustrated a simple set of equivalence rules as a first step towards identifying and removing differences in expression, but expanded work is required to provide a full treatment.

Following this more complete definition of equivalence (and potentially other types of) rules to identify and remove differences in expression, the work will also be extended to examine in more detail the actual differences in semantics that can be drawn from the NSM definitions, and how those differences can be used in GIS to assist in querying and automated processing.

## 7. Conclusions

The work presented in this paper is an attempt to allow non-expert users to easily define their own individual informal semantics in a way that can potentially be used in a geographic information system. This work, together with previous (Stock 2008) and ongoing work by the author indicates the promise of the semantic primitives for allowing comparison of meaning across individual and variable conceptualizations. Ultimately, it is hoped that the development of this approach may support the incorporation of informal, individual semantics into GIS, allowing systems that can accommodate context-sensitive and cultural variations in the definition of and interaction with geographic concepts.

## References

Bennett B, Mallenby D and Third A (2008) An Ontology for Grounding Vague Geographic Terms. In Eschenbach, C. and Gruninger, M. (eds) Proceedings of the 5th International Conference on Formal Ontology in Information Systems (FOIS-08). IOS Press, Amsterdam.

Cohn A, Bennett B, Gooday J and Gotts NM (1997) Qualitative Spatial Representation and Reasoning with the Region Connection Calculus *Geoinformatica* 1 pp275—316

Egenhofer M, Mark D and Herring J (1994) *The 9-Intersection: Formalism and Its Use for Natural-Language Spatial Predicates*. National Center for Geographic Information and Analysis, Santa Barbara.

Goddard C (1998) Bad arguments against semantic primitives *Theoretical Linguistics* 24 pp129-156

Dilo A, de By RA and Stein A (2007) A System Of Types And Operators For Handling Vague Spatial Objects *International Journal of Geographical Information Science* 21 pp397—426

Goddard C and Wierzbicka A (2002) Semantic Primes and Universal Grammar. In: Goddard, C., Wierzbicka, A. (eds.) *Meaning and Universal Grammar - Theory and Empirical Findings* (2 volumes). John Benjamins, Amsterdam.

Goddard C (2002) On and on: Verbal Explications for a Polysemic Network *Cognitive Linguistics* 13 pp277-294

Mark D (1993) Toward A Theoretical Framework For Geographic Entity Types. In Frank, A.U. and Campari, I. (eds) *Spatial Information Theory: A Theoretical Basis for GIS*, LNCS vol 716, pp 270-283. Springer-Verlag, Berlin.

Mark D and Frank A (1996) Experiential and Formal Models of Geographic Space *Environment and Planning B: Planning and Design* 23 pp3—24

Montello D, Goodchild M, Gottsegen J and Fohl P (2003) An Ontology for Grounding Vague Geographic Terms *Spatial Cognition and Computation* 3 pp185—204

Robinson V (1990) Interactive Machine Acquisition of Fuzzy Spatial Relation *Computers and Geosciences* 16 pp857—872

Robinson V (2000) Individual And Multi-personal Fuzzy Spatial Relations Acquired Using Machine-Human Interaction *Fuzzy Sets and Systems* 113 pp133—145

Schockaert S, Cornells C, De Cock M and Kerre EE (2006) Fuzzy Spatial Relations between Vague Regions. In: 2006 3rd International IEEE Conference on Intelligent Systems pp. 221—226. London.

Stock K (2008) Determining Semantic Similarity of Behaviour Using Natural Semantic Metalanguage to Match User Objectives to Available Web Services *Transactions in GIS* 12(6) pp733-755

Waters T and Evans A (2003) Tools for the Web-based GIS Mapping of Fuzzy Vernacular Geography. Paper presented at GISRUK 2003, 9-11 April 2003, London.

**Biography**

Dr Kristin Stock is a Research Fellow in geospatial semantics at the Centre for Geospatial Science, University of Nottingham. She undertakes and leads research in natural language and multi-lingual geospatial querying; the semantics of geo-social concepts and semantic knowledge infrastructures. Kristin also runs a consultancy specialising in geographic information interoperability.

# GIS analysis of historical marine geomorphology: the outer Thames seabed

H. Burningham and J.R. French

Coastal and Estuarine Research Unit, Department of Geography, University College London, Gower Street, London WC1E 6BT, UK
Tel: 020 7679 0577 | Email: h.burningham@ucl.ac.uk

Keywords: bathymetric change, morphodynamics

The outer Thames seabed comprises a series of tidal-stream aligned sandbanks, ridges and channels, the orientation of which are dependent on the proximity to the Thames estuary and coastal headlands such as Orfordness (Figure 1). This paper presents a 140 year history of changes in this seabed geomorphology across a region from Aldeburgh (Suffolk), to Southend-on-Sea (Essex) and Margate (Kent).

In this analysis, five hydrographic charts published from the mid-1800s to the present, at around 40-50 year intervals, were used to reconstruct historical changes in seabed morphology. Charts were georectified onto a common coordinate system (British National Grid), and soundings were digitised and converted to Ordnance Datum (OD) for comparative analysis. There are significant variations in tide range (and hence Chart Datum (CD)) across the region, which preclude the use of a constant vertical offset between CD and OD. A spatially-variable conversion was therefore devised (Figure 1). Depth data relative to OD were interpolated onto regular grids (3D surfaces) to allow spatial analysis of seabed evolution. Data preparation and analysis was conducted in ArcGIS and Matlab.

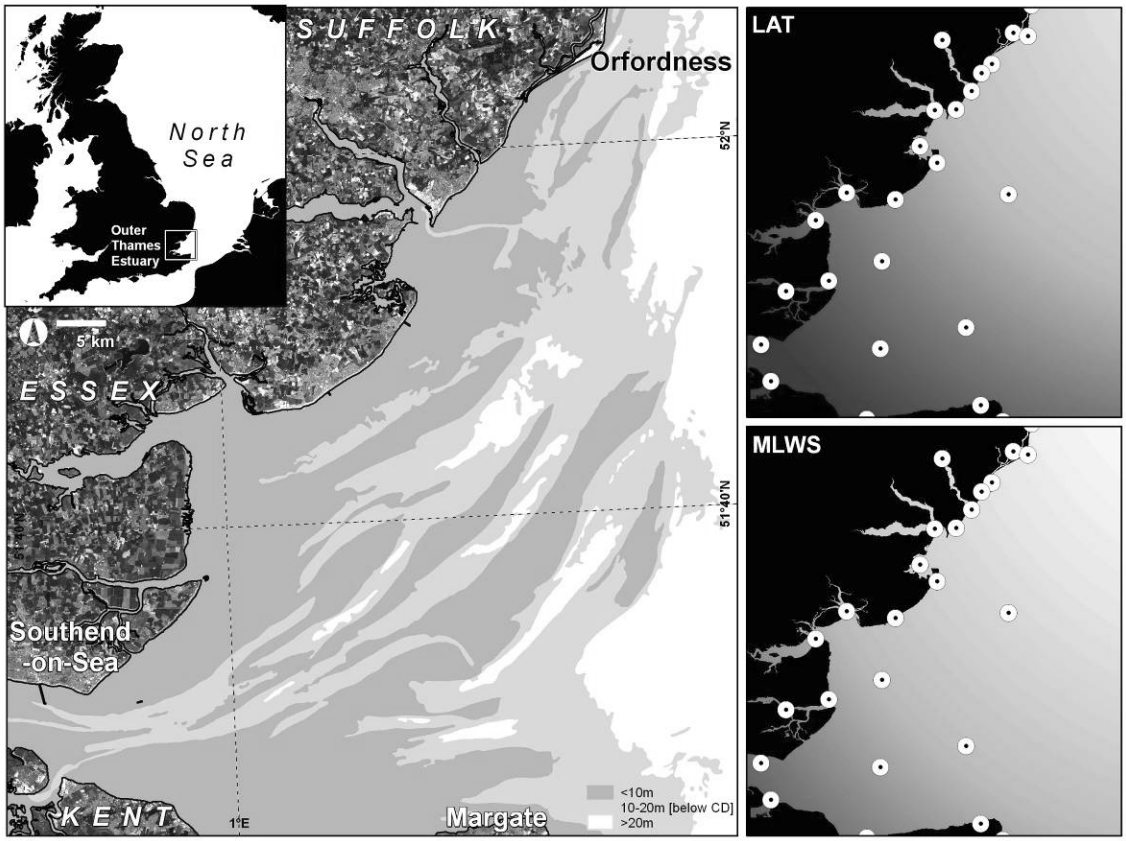



**Figure 1** Bathymetry of the outer Thames and trend surfaces for conversion of depth data from chart datum (e.g. LAT (lowest astronomical tide) and MLWS (mean low water springs)) to Ordnance Datum (OD).

Bathymetric analysis included computation of spatial statistics and net changes in seabed depth surfaces over the 140 year period to assess historical seabed behaviour, focusing on the analysis and significance of persistent trends. Descriptive statistics of the historical seabed depth surfaces reveal a number of key characteristics (Figure 2). First, the 140 year mean depth is broadly aligned to the current configuration of banks and channels. This illustrates the relative consistency in the historical organisation of seabed features, which are dominated by a) long shallow banks and deep intervening channels in the central part of the region and b) thin near-linear hair-pin ridges to the north of the region, off the Suffolk coast.

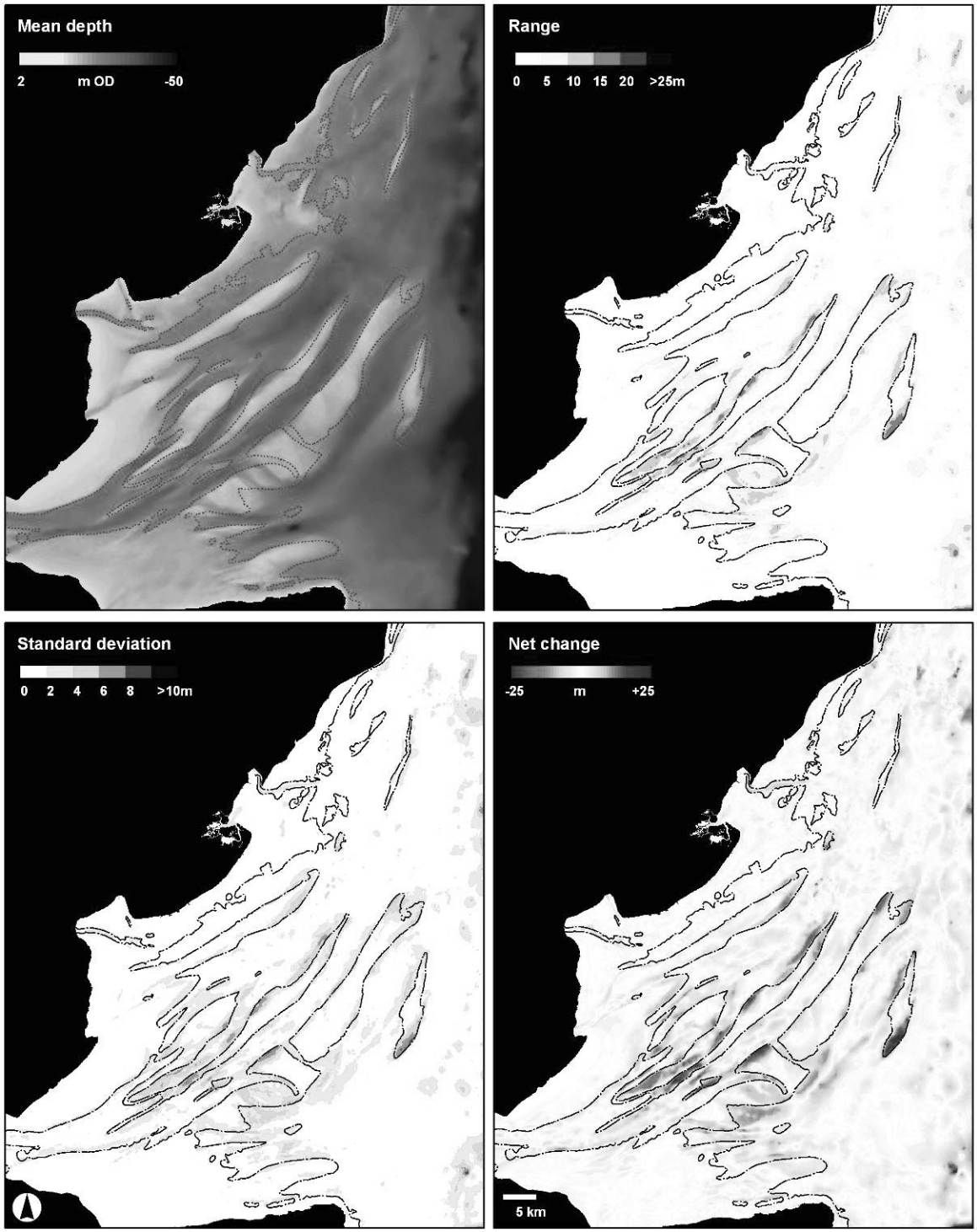



**Figure 2** Spatial analysis of bathymetric change across the outer Thames seabed between the 1860s and present (2000s). The 10m OD contour is shown for reference.

Second, the greatest variability in seabed depths is associated with bank margins. This is a clear indication that historical changes in depth are primarily associated with lateral shifts in bank and channel features. The final observation from the descriptive statistics is that net changes in seabed depths are equivalent to gross changes, demonstrating that the present seabed morphology is distinctly different to that in the mid-1800s.

Bathymetric change analysis centred on computation of the linear rate of change over the 140 year period. Over this spatial and temporal scale, there might be some expectation that sea-level forcing be evident in the morphological changes observed. Historical sea-level rise, estimated at 1.9-2.4 mm $yr^{-1}$ in this region equates to a net rise in sea level of around 27-35 cm, which is well within the error of the datasets and subsequent calculations. Van der Wal & Pye (2003) estimated a confidence interval of ± 0.58 m in their analysis of bathymetric change in the Ribble estuary (northwest England), taking into account changes in method of depth measurement and datum conversions. For the Thames, the inclusion of a modelled CD to OD conversion factor introduces further uncertainty (RMS ~0.25 m) into the analysis, meaning that changes of less than 1 m over the 140 year period (~7mm $yr^{-1}$) are unlikely to be statistically significant. Where seabed morphodynamics are dominated by lateral shifts in bank position, mirrored linear trends are evident along opposite margins (Figure 3). Statistically significant trends are present across the region, but are most notable around the central banks. Kentish Knock is progressively changing shape, and in the process, extending southward. Erosion along its northwest margin is mirrored by accretion to the south. Accretion continues along a southwesterly arm that appears to be connecting Kentish Knock to the southern banks of Long Sand. This is accompanied by a progressive deepening of the channel along the eastern margin of Long Sand.

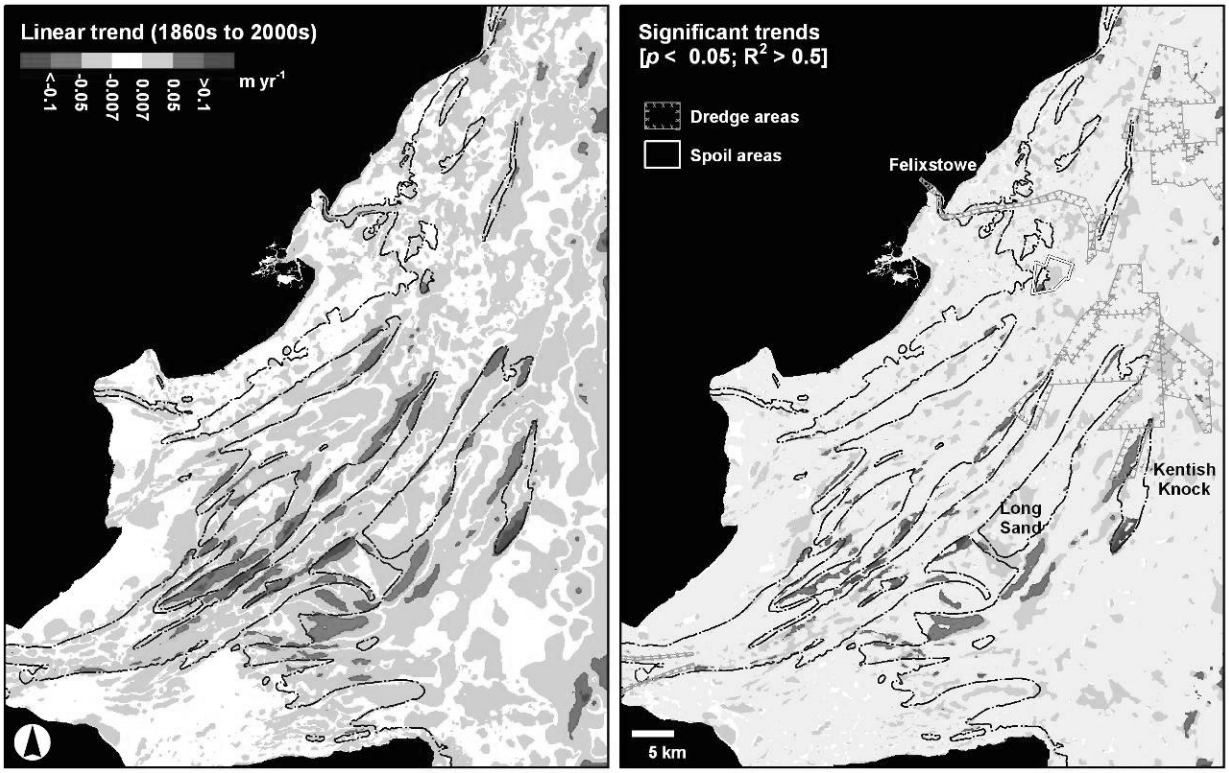



**Figure 3** Spatial analysis of linear trends in seabed change over the last 140 years. Linear trends are categorised into significant ($p < 0.05$ and $R^2 > 0.5$) and non-significant (blanked with light-grey mask in figure).

Whilst it is clear that vertical accretional and erosional changes across much of the broader seabed are not statistically significant, this analysis nevertheless reveals some more localised changes that are of interest . For example, regional growth to the southwest of Kentish Knock appears to be linked to the ongoing re-shaping of this bank. To the east of the northern tip of Long Sand, a relatively broad area of erosion coincides with dredging areas, but equally other dredge areas in the region appear to show accretion. Dredge material disposal grounds (spoil areas) appear to be more obviously associated with significant seabed change. In particular, South East Spit to the southeast of Felixstowe has experienced marked historical accretion, resulting in the seaward extension of the nearshore shelf along this section of coastline.

Elsewhere, patterns of change are largely related to the movement of bank margins, particularly in the central region dominated by long channels and banks. To the east and northeast of Felixstowe however, three narrow ridges lying near-parallel to the shoreline appear to suggest a broader reorganisation and reshaping of the larger subtidal landforms. Over the historical time-scale, the outermost and middle ridges have become increasingly connected by a hairpin bank joining their northern extremities. Ridge reorganisation and growth here appears to be fed by sediment scoured from spatially variable denudation of the seabed between these ridges. This may reflect changes in the position and/or strength of tidal-streams in this region, possibly in response to the extensive lateral shifts closer to the Thames. The tidal prism of the inner tidal-Thames is almost 1.5 times the combined prism of all other estuaries that link into this outer Thames system, so it is likely that changes in tidal forcing associated with the inner tidal-Thames could influence much of the long-term seabed dynamics across this region.

It is clear from the analysis undertaken here that there is considerable variability in the morphodynamic behaviour of the outer Thames seabed, both in the spatial extent of behaviours and the temporal-scale of trends. There is no conclusive evidence to suggest that sea-level rise is the dominant forcing mechanism, and shifting tidal streams seems may be a more likely cause of observed seabed dynamics, particularly in the vicinity of the major bank systems. Dredging and dredge material disposal are evident from the historic bathymetries, but these changes appear to be far less significant than those attributed to lateral channel and bank mobility.

## References

van der Wal, D. & Pye, K. 2003. The use of historical bathymetric charts in a GIS to assess morphological change in estuaries. *The Geographical Journal* **169** (1): 21-31.

# An assessment of three service area models for healthcare service analysis

Daniel Lewis[1]

[1]Department of Geography, Pearson Building, Gower Street, University College London, London, WC1E 6BT
Email: d.lewis@ucl.ac.uk | http://www.danieljlewis.org

KEYWORDS: service area modelling, spatial dimensions of healthcare, Southwark

## 1. Introduction

Recently, attention in primary healthcare has focused on the catchment area, specifically the announcement by Health Secretary Andrew Burnham (2009) that within a year general practice (GP) boundaries are to be abolished. This decision fits with the NHS general drive towards opening up choice in primary care (DoH, 2009). In explaining the move away from GP catchment areas, Burnham links provision of primary care to the postcode lottery in health (cf. Bungay, 2005; Lyon et al, 2004) stating that it will allow patients to register with GPs they "are currently excluded [from] by dint of their postcode" (Burnham, 2009).

It should be established at this point that a GP catchment area is an area defined by a GP within which home visits must be made by the GP if required, and within which a GP must accept a patient unless the GP has declared their patient list closed. The GP has discretion on whether to accept a patient who lives outside of the GP's catchment area. The GP has a responsibility to look after the health of the local community, and the catchment area has to be agreed with the appropriate Primary Care Trust (PCT) who manage primary care for a set area, as such a catchment area must to some extent reflect the local situation. This is a difficult measure however as the concept of 'local community' is not fixed; the now-defunct Commission on Integration and Cohesion used an area “within 15/20 minutes walking distance” (CoIC, 2007 p. 20) as shorthand for a community. It is clear from an investigation of patient registration data there exists a unique 'core' area that defines uptake for each GP surgery, and thus patients tend to treat primary care as a 'local' service.

One approach to defining a local area comes through analysing the spatial extent of GP patient registers. If locality is defined as being subject to the area bounded to include a given percentage of a service's users then a locality can be defined. The creation of bounded areas that define a service area is a practice in spatial analysis with several distinct branches, this paper focuses on service areas defined through point pattern analyses. Harris and Johnston (2008) have shown for schools that service area analyses can be useful in understanding whether a service is representative of the local area- something that will constitute an important finding in healthcare as well. Such analyses will be important in understanding the needs of GPs in the future if catchment areas are to be abolished; there is a possibility that some GPs will diverge from the local situation as they specialise in certain types of patients. This also raises larger questions related to segregation and post-residential sorting of service users (cf. Harris et al, 2007).

Previous work (Lewis, 2009) has used linear programming to create service areas based on accessibility to consider the role of accessibility and choice in primary care (cf. Allen, 2007 for a similar approach to schools). This paper will examine three methods of creating service areas and evaluate their suitability in healthcare research. The three methods detailed below derive from Huff and Batsell (1977), Gibin et al (2007) and Duckham et al (2008).

## 2. Three methods for modeling service areas

In the case of the Huff model (1977) and the Duckham model (2008) the service area is based on the network distance of patients from each GP. The Gibin (2007) model works with the density distribution of the patients as opposed to their distance from a surgery. In fact, each GP exhibits a strong distance decay effect with the highest concentrations of patients being closest to their respective GP.

The data used in this paper is the full 2009 patient register for Southwark PCT geocoded to postcode level. The register covers all patients living within Southwark and surrounding boroughs that use one of 47 Southwark GPs. This accounts for approximately 300,000 people.

### 2.1 The Huff model

The Huff and Batsell (1977) service area is specified by extracting from a point distribution the most extreme points from a facility and fitting cubic splines so that the area within these points is enclosed. The rationale behind fitting cubic splines given by Huff and Batsell (1977) is that they are more effective when dealing with sample data, as the data used in this paper is for the whole population it should be acceptable to enclose the service area with straight lines. Both cubic spline and straight lines were fitted for the purpose of the analysis.

The most distant points for a facility are extracted by rotating a 22.5$^{o}$ segment around the facility, performing a point-in-triangle operation for the points that are contained and recording the most distant point from the facility. The segment is incremented around the facility by 1$^{o}$ giving 360 points, some of which will overlap each other. A cubic spline is then fitted, or the points are simply joined together by lines (more information in Huff and Batsell, 1977 p. 582). Figure 1 shows an example of the output of the algorithm.

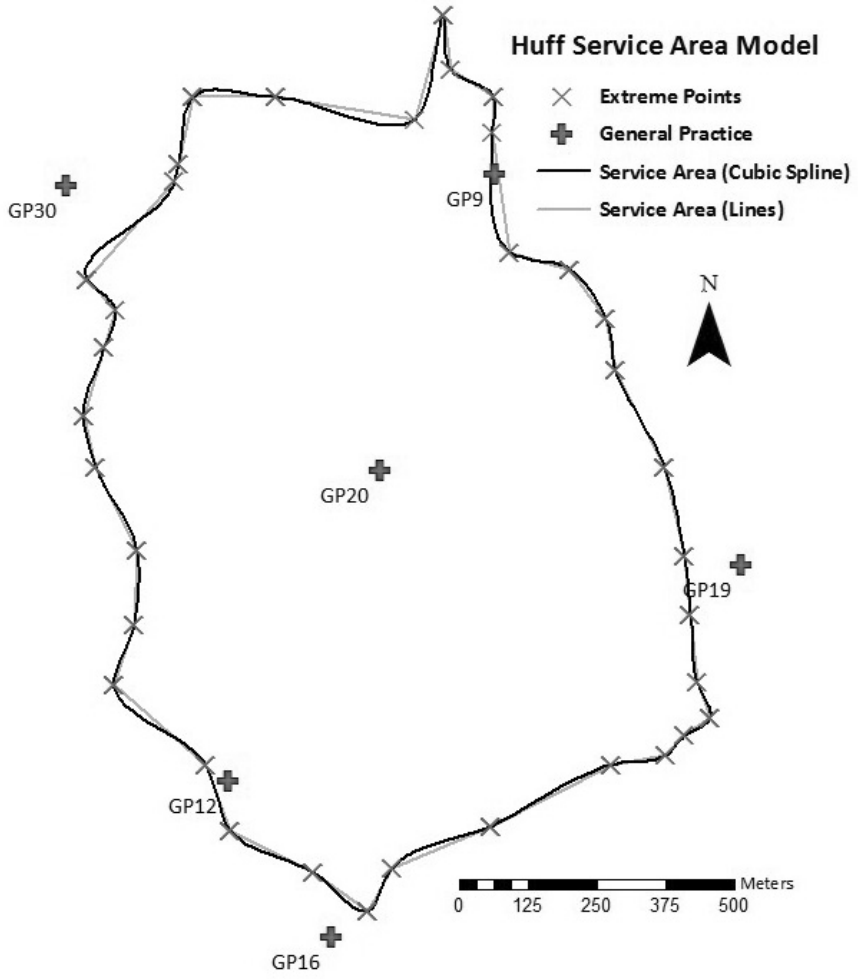



**Figure 1:** Huff Service Area derived at 60% of patient list for GP20.

### 2.2 The Duckham χ - shape model

This is not specifically a model for deriving service areas, however its usefulness is evident. Basic service area analyses have used the convex hull of a given point distribution. A convex polygon that completely encloses a set of points (de Smith et al, 2009) is known as the convex hull, these polygons can be subject to distortions by outlying points. One way of creating a more effective bounding shape for a set of points is to calculate the χ - shape (chi-shape) as defined by Duckham et al (2008).

Calculating the χ - shape involves first calculating the delaunay triangulation for a set of points; the delaunay triangulation is a complete set of non-overlapping triangles in which the outer boundary is the convex hull of the point set (de Smith et al, 2009). From this triangulation a normalised length value is computed and any boundary edge that is longer that this value is removed. Each time a boundary edge is removed a new set of boundary edges is computed and another edge removed until all boundary edges that are greater than the normalised length have been removed, or cannot be removed as they would leave a dangling line (more information in Duckham et al, 2008 p. 3226-9). Figure 2 shows an example output of the algorithm, the original and resultant delaunay triangulations are shown. The edges of the triangulation are found by attempting to navigate around 2 different triangles from each edge, if only 1 can be navigated the edge is a boundary edge.

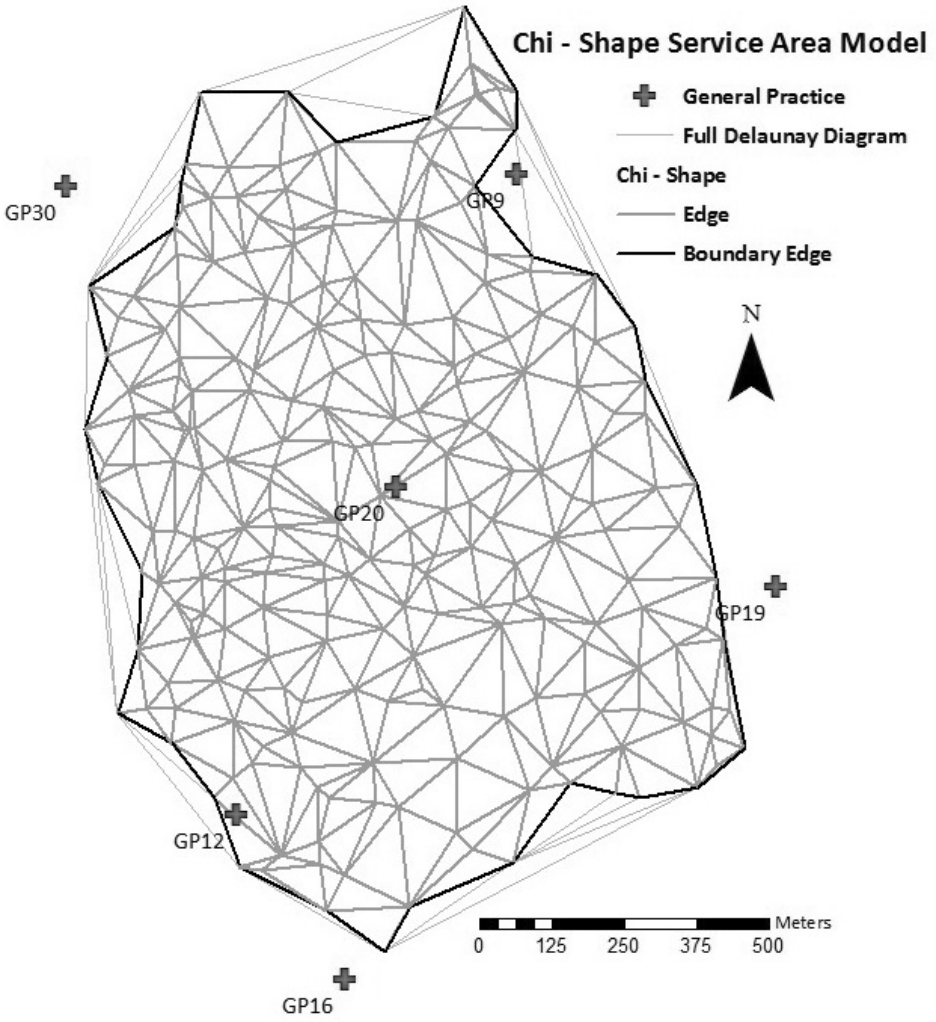



**Figure 2**: Duckham Chi - Shape Service Area derived at 60% of patient list for GP20.

### 2.3 The Gibin Model

The service area model used in Gibin et al (2007) derives an areal extent from a density surface specified to enclose a given percentage of the distribution. This method allows the possibility of deriving a spatial discontinuous service area for a facility, whereas the Huff and Duckham models assume that the service area will be singular, the Huff model also presupposes that the service area will include the facility.

A density surface is created using kernel density estimation (KDE) which creates a raster dataset in which each cell has a density value that is weighted according to the cell's distance from points in the point distribution. KDE requires the user to set a bandwidth (500m in Figure 3) and an output cell size (10m in Figure 3), different parameter values can create different results, thus specification is important. Having obtained a KDE surface for a given distribution, a PVC is calculated, a PVC is not a simple contour line connecting cells with the same value. Instead, the PVC delineated a region of the density surface that accounts for a given percentage of surface and hence a given percentage of the underlying point distribution. Figure 3 shows an output KDE with a PVC enclosing a given percentage of it.

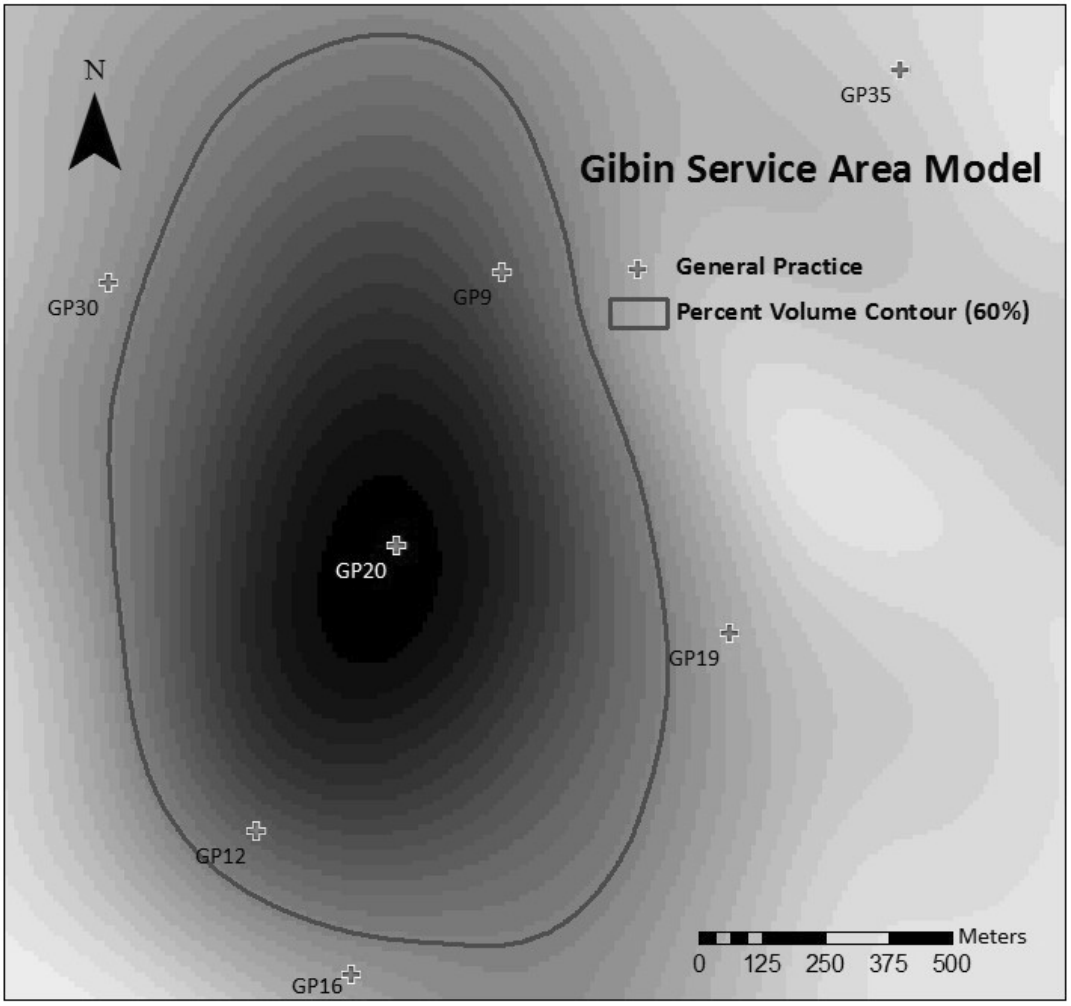



**Figure 3**: Gibin PVC service area model derived at 60% of patient list for GP20.

### 3. Analysis of service area models

Table 1 shows summary values for 47 GP service areas in Southwark calculated using each model (the full table is available as Appendix 1). Each service area was parameterised to enclose 60% of each GPs patient list. The table reports how successful this was, as well as some other attributes related to the complexity of the service area: area, perimeter and shape (calculated by the isoperimetric quotient Q (cf. Osserman, 1978) as per Equation 1).

$$Q = \frac{4\pi A}{P^2} \quad (1)$$

Where $A$ is the area of the shape and $P$ the perimeter. Q is a value that is 1 for a perfect circle, decreasing toward 0 as the shape becomes less circular.

**Table 1:** Summary of service areas using 3 models for 47 Southwark GPs

| Service Area Model | Mean Area ($km^2$) | Mean Perimeter (m) | Mean Shape | Mean % Pop Enclosed | Std Dev % Pop Enclosed |
|---|---|---|---|---|---|
| Huff Model (Lines) | 2.11 | 6.01 | 0.60 | 60.99 | 2.37 |
| Huff Model (Cubic Splines) | 2.15 | 6.02 | 0.59 | 58.32 | 3.96 |
| Duckham Chi-Shape Model | 1.99 | 6.40 | 0.53 | 60.87 | 1.64 |
| Gibin PVC Model | 1.38 | 4.67 | 0.83 | 67.55 | 3.91 |

In terms of enclosing the desired percentage of patients within a boundary, the Duckham model and Huff model fitted with lines rather than cubic splines are the most effective. The Duckham model also exhibits the lowest standard deviation suggesting stability across the set of GPs. The Gibin Model was notably the poorest, the high mean value for shape as well as low mean area and perimeter suggest the shapes created are the most generalised representations of the models tested. The performance is to some extent a result of the balancing act involved in specifying the bandwidth and cell size which effect the processing time and storage required. In this test the Gibin model took substantially longer to compute than the other models.

## 4. Conclusions and further work

The Huff lines and Duckham models seem the most appropriate for statistical analysis of local GP communities and patient - GP compositions. The Gibin and Huff Cubic Spline models may be more appropriate for visualisation as demonstrated by Gibin (2009) in the London Profiler website (www.londonprofiler.org). The isoperimetric quotient is useful in characterising the complexity of the resultant catchment, which may give an insight into the environmental, or social constraints on a particular GP's catchment. Particularly in the case of overlapping catchment areas which are a likelihood in the densely populated urban environment. Such overlapping catchments present an interesting challenge for the application of models of choice in this context. The Gibin model may be further improved using a data-based bandwidth which would vary for each GP, rather than simply specifying a common bandwidth as is currently the case.

These models use postcode georeferenced data, results may improve with data geocoded at the address level. This will allow for a more complete analysis of healthcare decisions locally as well as of segregation characteristics at the household and street level. Such analyses will allow an understanding of how choice-based policies actually sit in a societal context, and whether there exist any inequities in choice in healthcare. There is also scope to investigate these issues at a larger scale using a model of spatial interaction as per Wilson (2000).

## References

Allen R (2007) Allocating Pupils to Their Nearest Secondary School: The Consequences for Social and Ability Stratification *Urban Studies* 44(4) pp751-770.

Bungay H (2005) Cancer and Health Policy: The Postcode Lottery of Care *Social Policy and Administration* 39 pp35-48.

Burnham A (2009) Speech by the Rt Hon Andy Burnham, Secretary of State for Health, 17 September 2009 to the Kings Fund. Available at: http://www.dh.gov.uk/en/News/Speeches/DH_105366 [Accessed December 1, 2009]

CoIC (2007) *Commission on Integration and Cohesion - Our shared future* TSO, London

DoH (2009) *The NHS Constitution: the NHS belongs to us all* Department of Health

de Smith M, Goodchild M and Longley P (2009) *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools 3rd Edition* Winchelsea Press

Duckham M et al (2008) Efficient generation of simple polygons for characterizing the shape of a set of points in the plane *Pattern Recognition* 41(10) pp3224-3236

Gibin M, Longley P and Atkinson P (2007) Kernel Density Estimation and Percent Volume Contours in General Practice Catchment Area Analysis in Urban Areas *GISRUK 2007 Proceedings* Available at: http://ncg.nuim.ie/gisruk/materials/proceedings/PDF/5A3.pdf

Harris R, Johnston R and Burgess S (2007) Neighborhoods, Ethnicity and School Choice: Developing a Statistical Framework for Geodemographic Analysis *Population Research and Policy Review* 26(5) pp553-579

Harris R and Johnston R (2008) Primary Schools, Markets and Choice: Studying Polarization and the Core Catchment Areas of Schools *Applied Spatial Analysis and Policy* 1(1) pp59-84

Huff DL and Batsell RR (1977) Delimiting the Areal Extent of a Market Area *Journal of Marketing Research* 14(4) pp581-585

Lewis D, Mateos P and Longley P (2009) Choice and the composition of general practice patient registers *CASA Working Paper* 150 Available at: http://www.casa.ucl.ac.uk/working_papers/paper150.pdf

Lyon RM et al (2004) Surviving out of hospital cardiac arrest at home: a postcode lottery? *Emergency Medicine Journal* 21 pp619-624

Osserman R (1978) The Isoperimetric Inequality. *Bulletin of the American Mathematical Society* 84(6) pp1182-1238.

Wilson AG (2000) Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis Prentice Hall, London

**Biography**

*Daniel Lewis is a 2nd year PhD research student in the Department of Geography, University College London. He has a BA in Geography from LSE, and an MSc in GIS from UCL, his PhD is funded by an ESRC CASE award sponsored by Southwark Primary Trust.*

| | Huff Model (Lines) | | | | Huff Model (Cubic Splines) | | | | Duckham Chi - Shape Model | | | | Gibin Density model | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GP | Area | Perimeter | Shape | % Pop Enclosed | Area | Perimeter | Shape | % Pop Enclosed | Area | Perimeter | Shape | % Pop Enclosed | Area | Perimeter | Shape | % Pop Enclosed |
| 1.00 | 2.04 | 6.84 | 0.55 | 59.30 | 2.07 | 6.94 | 0.54 | 57.87 | 2.04 | 7.03 | 0.52 | 60.55 | 2.31 | 8.62 | 0.39 | 63.59 |
| 2.00 | 1.94 | 5.69 | 0.75 | 60.60 | 1.95 | 5.76 | 0.74 | 60.17 | 1.96 | 5.66 | 0.77 | 61.11 | 1.76 | 4.98 | 0.89 | 63.93 |
| 3.00 | 1.10 | 6.00 | 0.39 | 63.54 | 1.11 | 6.14 | 0.37 | 61.15 | 0.80 | 7.25 | 0.19 | 61.66 | 1.09 | 3.93 | 0.89 | 65.58 |
| 4.00 | 1.12 | 4.71 | 0.63 | 58.77 | 1.12 | 4.91 | 0.59 | 59.12 | 1.16 | 4.70 | 0.66 | 61.36 | 0.95 | 4.01 | 0.74 | 73.64 |
| 5.00 | 0.43 | 3.62 | 0.41 | 58.52 | 0.44 | 3.74 | 0.40 | 44.54 | 0.29 | 5.45 | 0.12 | 59.07 | 0.79 | 3.39 | 0.86 | 75.13 |
| 6.00 | 0.96 | 5.00 | 0.48 | 62.67 | 0.96 | 4.75 | 0.53 | 61.52 | 0.95 | 4.95 | 0.49 | 62.45 | 1.09 | 3.72 | 0.99 | 64.22 |
| 7.00 | 0.92 | 4.81 | 0.50 | 62.38 | 0.93 | 4.68 | 0.53 | 61.93 | 0.88 | 5.18 | 0.41 | 62.78 | 0.85 | 3.48 | 0.89 | 77.41 |
| 8.00 | 0.26 | 2.51 | 0.52 | 57.26 | 0.27 | 2.64 | 0.48 | 54.86 | 0.22 | 3.14 | 0.28 | 61.65 | 0.48 | 2.46 | 0.99 | 74.10 |
| 9.00 | 1.98 | 6.52 | 0.59 | 60.95 | 2.01 | 6.69 | 0.56 | 60.27 | 1.83 | 6.82 | 0.49 | 61.29 | 2.11 | 7.12 | 0.52 | 66.64 |
| 10.00 | 1.51 | 5.11 | 0.73 | 62.92 | 1.52 | 5.14 | 0.72 | 60.96 | 1.52 | 5.09 | 0.74 | 62.86 | 1.35 | 4.25 | 0.94 | 62.46 |
| 11.00 | 0.39 | 3.20 | 0.48 | 59.76 | 0.41 | 3.25 | 0.48 | 54.02 | 0.40 | 3.20 | 0.49 | 60.34 | 0.57 | 2.76 | 0.95 | 75.23 |
| 12.00 | 1.47 | 5.36 | 0.64 | 62.72 | 1.49 | 5.51 | 0.62 | 61.43 | 1.55 | 5.04 | 0.77 | 61.23 | 1.62 | 5.30 | 0.72 | 65.52 |
| 13.00 | 0.70 | 4.28 | 0.48 | 63.86 | 0.72 | 4.21 | 0.51 | 66.19 | 0.69 | 4.36 | 0.46 | 62.21 | 0.77 | 3.14 | 0.98 | 70.86 |
| 14.00 | 1.28 | 4.54 | 0.78 | 60.11 | 1.29 | 4.60 | 0.77 | 56.84 | 1.22 | 4.88 | 0.64 | 58.97 | 1.52 | 4.41 | 0.99 | 63.32 |
| 15.00 | 0.28 | 3.09 | 0.38 | 57.12 | 0.28 | 3.15 | 0.35 | 51.17 | 0.32 | 2.82 | 0.50 | 57.88 | 0.67 | 2.99 | 0.95 | 72.21 |
| 16.00 | 1.96 | 6.94 | 0.51 | 60.87 | 2.05 | 7.09 | 0.51 | 56.21 | 2.17 | 6.36 | 0.67 | 61.18 | 2.14 | 6.68 | 0.60 | 63.00 |
| 17.00 | 1.33 | 5.39 | 0.57 | 59.48 | 1.34 | 5.55 | 0.55 | 56.60 | 1.26 | 5.53 | 0.52 | 62.47 | 1.12 | 4.18 | 0.80 | 68.87 |
| 18.00 | 0.85 | 3.91 | 0.70 | 61.68 | 0.85 | 3.98 | 0.68 | 59.51 | 0.82 | 4.06 | 0.63 | 60.77 | 1.04 | 3.75 | 0.93 | 71.60 |
| 19.00 | 2.65 | 6.70 | 0.74 | 59.69 | 2.68 | 6.79 | 0.73 | 58.29 | 2.67 | 6.64 | 0.76 | 59.96 | 2.28 | 6.06 | 0.78 | 66.02 |
| 20.00 | 1.18 | 4.62 | 0.69 | 60.17 | 1.18 | 4.51 | 0.73 | 57.39 | 1.15 | 4.75 | 0.64 | 57.18 | 1.56 | 4.71 | 0.88 | 65.32 |
| 21.00 | 0.90 | 3.70 | 0.82 | 59.89 | 0.91 | 3.76 | 0.81 | 56.87 | 0.90 | 3.74 | 0.80 | 60.05 | 1.07 | 3.79 | 0.94 | 69.25 |
| 22.00 | 0.94 | 5.20 | 0.44 | 57.17 | 0.96 | 5.34 | 0.42 | 53.78 | 0.86 | 5.56 | 0.35 | 56.35 | 1.34 | 5.21 | 0.62 | 71.55 |
| 23.00 | 2.62 | 7.10 | 0.65 | 61.22 | 2.61 | 7.22 | 0.63 | 59.41 | 2.67 | 6.87 | 0.71 | 61.57 | 2.62 | 6.22 | 0.85 | 64.57 |
| 24.00 | 0.49 | 3.90 | 0.41 | 57.84 | 0.49 | 3.85 | 0.41 | 49.32 | 0.42 | 4.34 | 0.28 | 60.05 | 0.86 | 3.42 | 0.92 | 64.28 |
| 25.00 | 3.95 | 8.11 | 0.76 | 61.77 | 3.97 | 8.25 | 0.73 | 60.57 | 4.05 | 7.85 | 0.83 | 62.34 | 1.13 | 3.74 | 0.88 | 67.31 |
| 26.00 | 2.21 | 5.66 | 0.86 | 63.47 | 2.23 | 5.76 | 0.84 | 62.81 | 2.04 | 6.11 | 0.69 | 61.11 | 1.67 | 8.87 | 0.27 | 65.27 |
| 27.00 | 1.41 | 5.90 | 0.51 | 60.93 | 1.45 | 5.99 | 0.51 | 55.70 | 1.36 | 5.68 | 0.53 | 60.93 | 1.52 | 4.78 | 0.83 | 67.30 |
| 28.00 | 3.38 | 8.25 | 0.62 | 59.11 | 3.40 | 8.44 | 0.60 | 57.36 | 3.15 | 9.42 | 0.45 | 60.08 | 3.37 | 10.00 | 0.42 | 64.15 |
| 29.00 | 1.71 | 5.58 | 0.69 | 61.23 | 1.71 | 5.56 | 0.70 | 58.53 | 1.59 | 6.20 | 0.52 | 57.88 | 1.58 | 4.70 | 0.90 | 65.29 |
| 30.00 | 0.68 | 4.46 | 0.43 | 59.87 | 0.00 | 0.18 | 0.46 | 55.69 | 0.80 | 3.89 | 0.66 | 63.67 | 1.04 | 3.80 | 0.90 | 68.78 |
| 31.00 | 1.45 | 5.63 | 0.57 | 60.91 | 1.44 | 5.41 | 0.62 | 59.58 | 1.36 | 6.13 | 0.46 | 59.01 | 1.42 | 4.31 | 0.96 | 63.85 |
| 32.00 | 1.07 | 4.10 | 0.80 | 61.49 | 1.08 | 4.14 | 0.79 | 59.02 | 1.00 | 4.42 | 0.64 | 60.92 | 1.07 | 3.76 | 0.95 | 68.01 |
| 33.00 | 0.74 | 4.42 | 0.48 | 58.07 | 0.76 | 4.62 | 0.45 | 57.47 | 0.75 | 4.43 | 0.48 | 59.35 | 1.16 | 3.90 | 0.96 | 68.45 |
| 34.00 | 0.46 | 2.91 | 0.68 | 57.51 | 0.46 | 3.04 | 0.62 | 56.17 | 0.48 | 2.88 | 0.73 | 59.68 | 0.72 | 3.05 | 0.97 | 69.02 |
| 35.00 | 0.73 | 3.54 | 0.73 | 60.26 | 0.73 | 3.70 | 0.67 | 53.60 | 0.66 | 3.90 | 0.55 | 61.17 | 1.06 | 4.49 | 0.66 | 67.52 |
| 36.00 | 1.10 | 4.46 | 0.70 | 60.05 | 1.11 | 4.59 | 0.66 | 58.11 | 1.05 | 4.76 | 0.58 | 59.38 | 1.26 | 4.04 | 0.97 | 66.36 |
| 37.00 | 7.44 | 13.85 | 0.49 | 60.12 | 7.62 | 14.26 | 0.47 | 58.41 | 4.97 | 17.39 | 0.21 | 60.43 | 3.21 | 9.54 | 0.44 | 62.25 |
| 38.00 | 7.82 | 12.78 | 0.60 | 64.95 | 7.95 | 13.11 | 0.58 | 61.21 | 7.17 | 13.09 | 0.53 | 62.99 | 1.23 | 3.99 | 0.97 | 65.34 |
| 39.00 | 0.57 | 4.36 | 0.38 | 64.39 | 0.58 | 4.40 | 0.38 | 60.90 | 0.51 | 4.55 | 0.31 | 62.73 | 0.54 | 2.62 | 0.99 | 72.60 |
| 40.00 | 3.47 | 8.94 | 0.55 | 66.51 | 3.53 | 9.16 | 0.53 | 65.48 | 2.80 | 11.50 | 0.27 | 62.85 | 0.55 | 2.64 | 1.00 | 70.90 |
| 41.00 | 12.70 | 16.74 | 0.57 | 61.79 | 13.27 | 17.24 | 0.56 | 58.96 | 11.73 | 17.39 | 0.49 | 63.07 | 1.78 | 5.52 | 0.73 | 64.60 |
| 42.00 | 8.90 | 13.98 | 0.57 | 68.65 | 10.01 | 14.23 | 0.62 | 66.92 | 9.05 | 13.44 | 0.63 | 61.19 | 1.08 | 3.72 | 0.98 | 66.87 |
| 43.00 | 4.27 | 8.98 | 0.66 | 62.44 | 4.21 | 9.24 | 0.62 | 60.59 | 4.75 | 13.06 | 0.35 | 60.46 | 1.61 | 4.76 | 0.89 | 63.41 |
| 44.00 | 0.52 | 3.21 | 0.63 | 62.21 | 0.52 | 3.30 | 0.60 | 58.20 | 0.52 | 3.25 | 0.62 | 63.62 | 0.73 | 3.06 | 0.97 | 73.16 |
| 45.00 | 1.16 | 6.01 | 0.40 | 61.05 | 1.17 | 6.15 | 0.39 | 58.25 | 1.06 | 6.42 | 0.32 | 61.78 | 1.26 | 4.55 | 0.76 | 68.71 |
| 46.00 | 2.36 | 6.46 | 0.71 | 60.78 | 2.39 | 6.66 | 0.68 | 58.70 | 2.27 | 6.79 | 0.62 | 60.81 | 2.32 | 5.90 | 0.84 | 63.89 |
| 47.00 | 1.58 | 5.18 | 0.74 | 60.43 | 1.60 | 5.23 | 0.74 | 59.18 | 1.62 | 5.03 | 0.81 | 60.71 | 1.78 | 5.10 | 0.86 | 63.40 |

# Automatic identification of High Streets and classification of Urban Land Use in Large Scale Topographic Database

Omair Z Chaudhry[1], Médéric Gravelle[2], Nicolas Regnauld[3]

[1] Manchester Metropolitan University,
Tel. +44 (0) 161247 1574, Department of Environmental and Geographical Science,
Manchester, M1 5GD
Email: O.Chaudhry@mmu.ac.uk

[2] Ecole Nationale des Sciences Géographiques, 77455 Marne la Vallée, France
[2] Ordnance Survey (GB), Romsey Road, Southampton, SO16 4GU

KEYWORDS: classification, land use, database enrichment, map generalisation

## 1. Introduction

This paper focuses on automatic detection of 'high streets' and classification of urban area into residential, commercial and industrial zones from a large scale topographic database. This research builds on past achievements in the field of map generalisation to perform this detection and classification process.

### 1.1 Background

National mapping agencies (NMAs) across the globe provide spatial datasets (and maps) at different levels of detail (or scale). Different scales reveal different information which is useful for different applications (Mackaness, 2007). It has been long desired that the process of transformation from detailed to generalised spatial datasets can be done automatically. This is the process of map generalisation. Map generalisation presents the required information at higher level of abstraction by automatically transforming large scale data.

### 1.2 Context and Objectives

Map generalisation is one of the most active areas of research at Ordnance Survey, the national mapping agency of Great Britain (GB). OS most detailed spatial product is called OS MasterMap®. Its Topography layer contains approximately half a billion real world features at a scale of 1:1250 in urban areas, 1:2,500 in rural areas and 1:10,000 in mountain and moorland areas. OS MasterMap also has a Address Layer which provides functional classification of each building feature in the Topography layer. It classifies buildings into house, flat, shop, warehouse, school, hospital, airport terminals. Such a detailed functional classification is quite useful for applications that require large scale datasets. But at lower levels of detail, such as 1:50,000 or 1:250,000, representation of such detailed information creates a cluttered output as shown in Figure 1. In Figure 1a buildings in OS MasterMap Topography layer are classified into residential, commercial or industrial classes using Address layer information. The level of information represented is quite appropriate for visualisation. On the other hand in Figure 1b the same information is difficult to visualise at 1:25,000 scale. Thus we need to generalise the data.

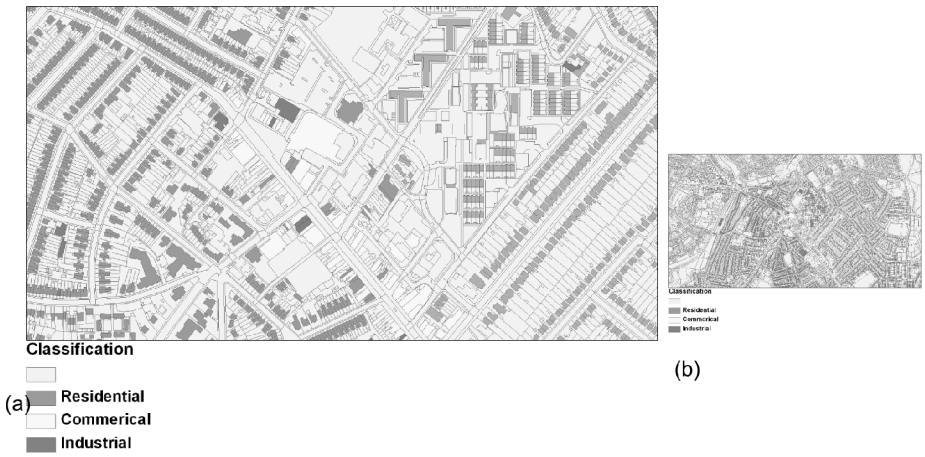



**Figure 1:** (a) Building classification at 1:1250 scale (b) representation at 1:25,000 scale gets cluttered thus in need of generalisation

In this paper we present an automatic technique for the classification of urban areas into residential, commercial and industrial blocks (section 2) so that this information can be properly represented at small scales. In section 3 we present the technique for automatic detection of commercial roads or high streets in urban areas. Such classification of roads is absent in the source dataset and here we present an automated solution and some results. Identification of such features is not only important from a generalisation perspective but also for understanding of urban morphology and urban economics (Hillier 1999; Lloyd et al 2003).

## 2. Urban block classification

### 2.1 Extraction of Urban Area

The first step for automatic classification of urban area, from a large scale source database (OS MasterMap), is to construct urban area. As stated earlier in our source dataset we store real world features such as buildings, road, pavements, land parcels but there is no representation of more abstract features such as urban area. Automated techniques have been proposed in research (Boffet, 2001; Small et al 2005, Chaudhry and Mackaness 2008) for derivation of urban extents. Chaudhry and Mackaness 2008 technique based on density of buildings was used for automatic extraction of urban areas from our source database. This approach calculates the density of each building using the total area and total distance of its (fifty) closest neighbours. This density value is then used to build the boundary by buffering and amalgamation of overlapping buffers. Figure 2 shows the boundary of city of Glasgow created via this approach.

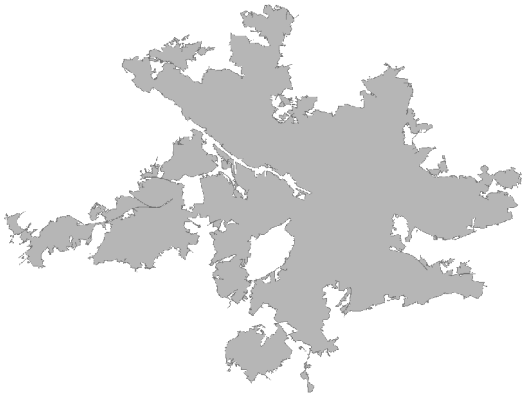


**Figure 2:** Automatically derived urban extent for city of Glasgow from source dataset using Chaudhry and Mackaness (2008)

## 2.2 Partition and Classification

The automatically derived urban areas are then partitioned into blocks using the road network data, Integrated Transport Network Layer (ITN). Each block can then be simply classified by simply counting the number of industrial, commercial and residential buildings contained by it. The highest of these can be assigned as the class of the block. Although straight forward this approach has a number of issues. Consider the illustration in Figure 3 we usually have large partitions at the periphery of an urban area. By just using the count as criteria the block in Figure 3 will be classified as residential whereas in reality residential area is only a fraction of the total area. Also this criterion of count ignores that industrial buildings are much larger in size as compared to residential buildings.

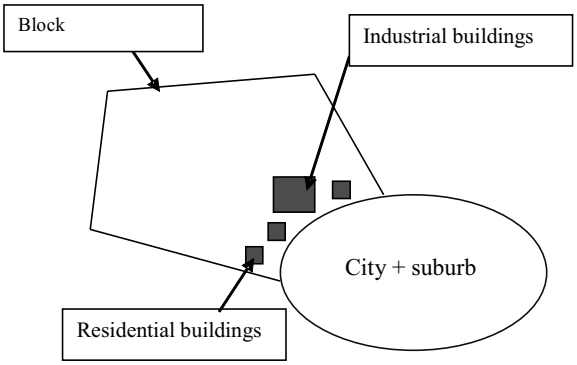



**Figure 3:** Classification of urban blocks at the edge of a city needs to take into account the size of block

In order to deal with such issues, instead of counting we calculated the ratio of total area of buildings of each type with the area of the block. The block is then classified by selecting the highest of the three (residential, commercial and industrial ratios) and also checking if the selected ratio is more than 5%. If not the block is not classified. Figure 4a presents the results of the urban block classification for the city of Glasgow. As observed many adjacent blocks have the same classification so it might be more appropriate for small scale representations that adjacent blocks having same classification are dissolved as shown in Figure 4b.

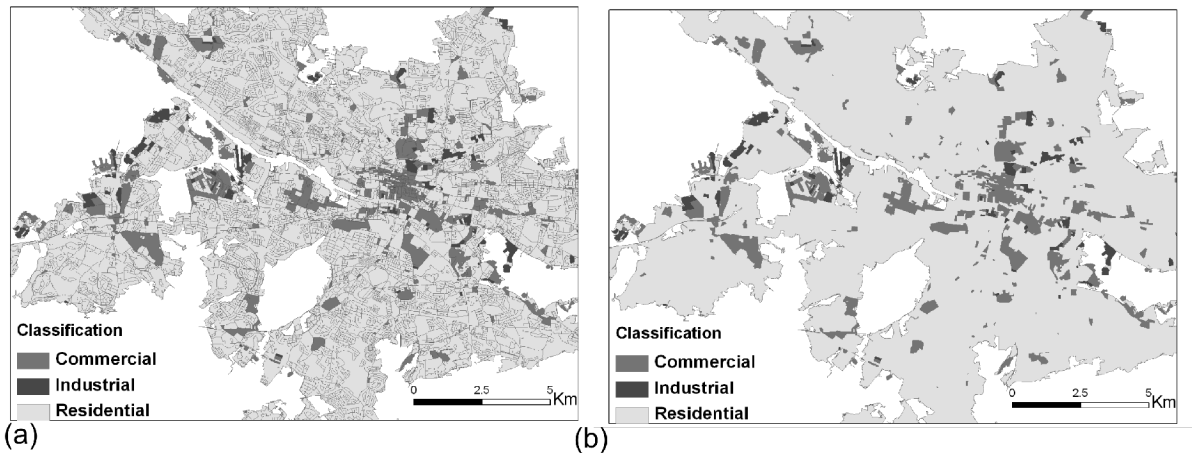



**Figure 4:** (a) Urban block classification for city of Glasgow (b) Aggregation of similar classification blocks

## 3. Extraction of High Streets

Typically in GB a high street or commercial road of a town or city is usually a long 'straight' street or road where there is a (usually) high density of retail. Usually they are often central to a town and

typically they have high pedestrian concentrations. OS MasterMap provides detailed road network of GB which also provides classification of roads into Motorways, A road, B road, local streets in its ITN Layer. But it does not classify any of these into high street or commercial roads.

Figure 5 illustrates a road immediately surrounded with commercial buildings which are followed by residential buildings on both sides. With such a block structure, the higher area ratio previously described is the residential one, and is more than 5%. So the block 1 and 2 are both classified as residential, and the commercial road will not be displayed. But such roads (high streets) are commonly represented at small scales and are frequently referred in everyday life. Here we present an approach for automatic identification of such streets.

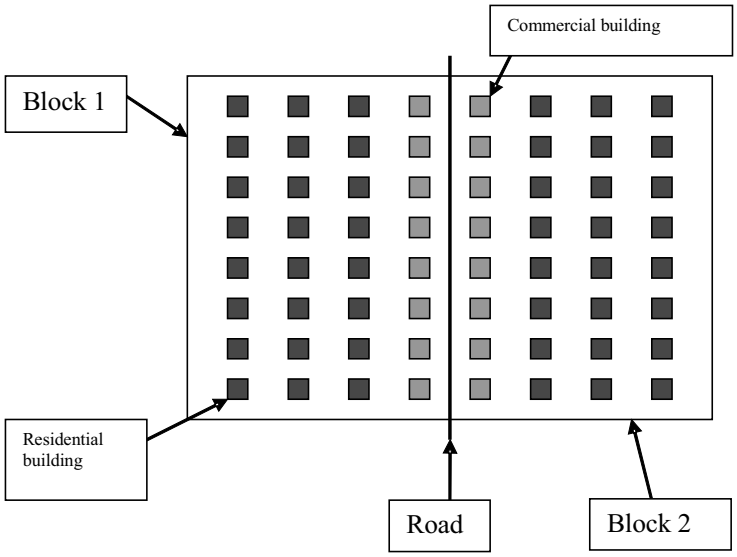



**Figure 5 :** Requirement of identification of commercial roads that are detected by the urban block classification

### 3.1 Clustering of commercial buildings

In order to extract commercial roads or high streets from our large scale database the first step was to carry out clustering. Clustering is one of the most well researched spatial data mining techniques (Andres and Sester, 2000). The most powerful methods of clustering in difficult problems, which give results having the best agreement with human performance, are the graph-based methods (Jaromczyk and Toussaint, 1992). A graph based clustering method involves computing a neighbourhood or proximity graph of the original dataset and then segmenting the graph into clusters by deleting any edge in the graph that is longer than some criterion.

In this research we have used a graph based approach for clustering, as described in Regnauld (2003) for supporting the amalgamation of buildings. We start with selection of commercial features from our source database (OS MasterMap Address layer). We then create a proximity graph using Delaunay triangulation. Using the length of the edges as criteria a MST (minimum spanning tree) is built. Clusters are then created by segmenting the MST - removing all the edges whose length is more than 100 metres. This process is illustrated in Figure 6

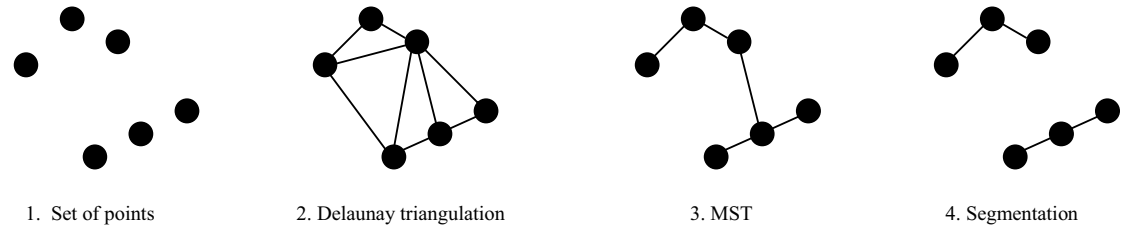



**Figure 6:** Steps for creating a cluster from source datasets points (OS Address Layer)

### 3.2 Constructing High Streets

The next step is to select road features from the source dataset (OS ITN) that are 'within' the clusters created in previous step. This is done by buffering each edge of each cluster by 50m and aggregating the overlapping buffers. Thus a one dimensional cluster is converted into a two dimensional polygon (Figure 7a) which is then used to select all road objects that lay inside.

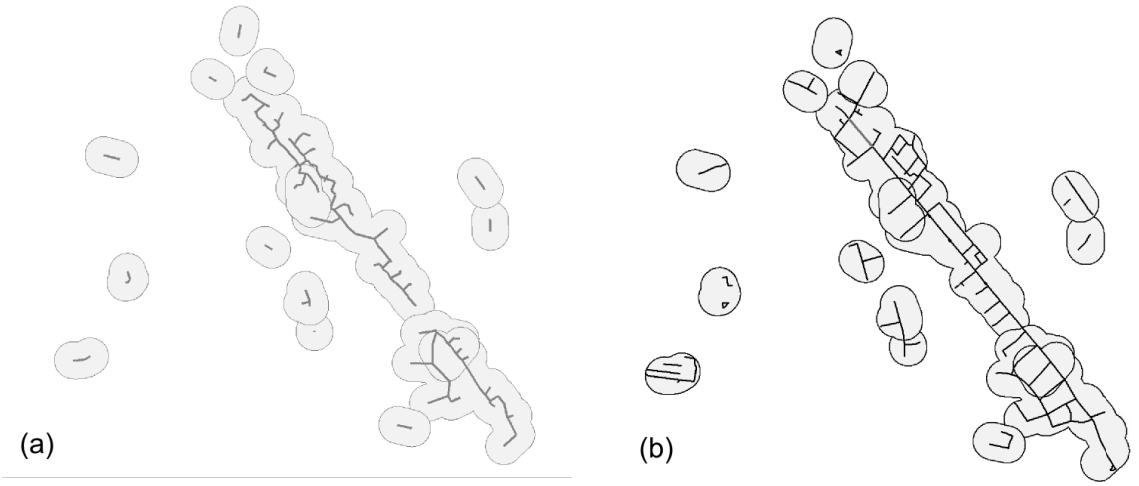



**Figure 7:** (a) Polygons created from clusters by buffering each edge of each cluster by 50m (b) road segments selected from cluster polygons (green). Highlighted segment illustrates that road objects are broken into segments at each intersection

The road network data in our source dataset (OS ITN) is stored as graph theoretic elements i.e. segments and nodes (Figure 7b). This means that roads are broken up at each junction into road segments. So in order to create a high street or long commercial road we need to combine these segments. For each selected road segment the neighbour roads are selected and the one that has the angle with the initial road closest to $\pi$ becomes part of the initial road. The process is then repeated for this newly added road segment. If the angle at any stage between initial road and neighbour is more than $3*\pi/4$ the neighbour road then becomes a new initial road. This process results in long 'straight' roads and is similar to the process of creating *strokes* as proposed in generalisation research (Thom, 2005; Thomson, and Brooks, 2000). Figure 8 illustrates the resulting road created for clusters shown in Figure 7. Since the road segments have been combined long straight roads (commercial roads) can now be selected using their lengths as highlighted in Figure 8.


**Figure 8:** Resultant strokes created from selected road segment. Symbol of each stroke is proportional to its length. The highlighted road is more than 1000m and is the Shirley high street in Southampton

## Conclusion

In this paper we presented how functional information can be represented at small scales for urban areas. So for the purpose of making maps, such functional information provides a good alternative level of information between the depiction of individual buildings and the depiction of just urban extents. We have also presented how functional information can be automatically derived using clustering and generalisation techniques for a new high level concept – high street. Such information can be even more valuable for applications that do perform spatial analysis. For example, these commercial areas can be useful for people planning bus routes, or for people looking for suitable areas for buying a new home.

## References

Anders, K.H & Sester, M (2000) Parameter-free cluster detection in spatial databases and its application to typification. In Comm. IV, ISPRS Congress, vol. 23, Part B4/1, pp.75-82, Amsterdam

Boffet, A (2001): *Méthode de création d'information multi-niveaux pour legénéralisation cartographique de l'urbain*. Université de Marne La Vallée, laboratoire COGIT, 2001.

Chaudhry, O. Z. & Mackaness, W.A. (2008) Automatic Identification of Urban Settlement Boundaries for Multiple Representation Databases. *Computer Environment and Urban Systems*. 32(2), pp. 95-109

Hillier (1999). Centrality as a process: accounting for attraction inequalities in deformed grids, Urban Design International 4, pp. 107–127.

Jaromczyk, J. and Toussaint, G., 1992. Relative neighborhood graphs and their relatives. In: Proceedings IEEE, Vol. 80(9 ), pp. 1502–1517.

Lloyd, D., Haklay, M., Thurstain-Goodwin, M., and Tobón, C. (2003) Visualising structure in new data systems. In: Longley, P. and Batty, M. (eds) Advanced spatial analysis. Redland, CA: ESRI Press, pp. 267–286.

Mackaness, W.A. (2007). Understanding Geographic Space. In Generalisation of Geographic Information: Cartographic Modelling and Applications (eds W.Mackaness, A. Ruas & L.T. Sarjakoski), pp. 1-10. Elsevier, Oxford

Regnauld, N. (2003). Algorithms for the amalgamation of topographic data. In: Proceedings of the 21st International Cartographic Conference, Durban, South Africa

Small, C., Pozzi, F., & Elvidge, C. D. (2005). Spatial analysis of global urban extent from DMSP-OLS night lights. Remote Sensing of Environment, 96, 277–291

Thom, S. (2005), A Strategy for collapsing OS Integrated Transport Network (ITN) dual carriageways, presented at the 8th ICA Workshop on Generalisation and Multiple Representation, La Coruña, Spain

Thomson, R. C., Brooks, R. (2007), Generalisation of Geographic Networks, In Generalisation of Geographic Information: Cartographic Modelling and Applications (eds W.Mackaness, A. Ruas & L.T. Sarjakoski), pp. 255-268. Elsevier, Oxford

# Supporting spatial negotiations in land use planning

Gustavo Arciniegas, Ron Janssen

Institute for Environmental Studies, Vrije Universiteit, De Boelelaan 1085 1081 HV Amsterdam
Tel. +31-20-5989 555 Fax +31-20-5989 553
Gustavo.Arciniegas@ivm.vu.nl, http://www.ivm.vu.nl

KEYWORDS: negotiation, participatory planning, spatial decision support, touch table, collaboration

## 1. Introduction

The integration of Multicriteria Analysis (MCA) and Geographic Information Systems (GIS) has contributed to the development of Spatial Decision Support Systems (SDSS) (Malczewski, 2006). While SDSS have mainly focused on individual decision-making, this focus has recently shifted to more participatory contexts, increasing the need to integrate SDSS with participatory approaches. Multiple stakeholders with conflicting objectives are characteristic of spatial decision problems (e.g. land use planning, environmental management). This paper describes a method that combines workshop meetings, GIS-MCA tools and an interactive mapping device called the 'Touch table' to support stakeholder negotiations for the participatory design of land use plans amidst conflicting objectives. The Touch table is a large touch screen that allows simultaneous input from up to four users. Section 2 lists recent approaches to support group-based spatial decision-making. Section 3 describes our workshop-based approach to support planning processes. Section 4 deals particularly with the steps of a negotiation workshop and briefly presents an implementation of it in a negotiation exercise with stakeholders involved in a case study area. Conclusions are presented in Section 5.

## 2. Supporting spatial negotiations

High levels of conflict and, therefore negotiation management, are important concerns within participatory spatial decision-making (Jankowski and Nyerges, 2001). Maps can also be a source of conflict in multi-actor planning, which can introduce further complexity to conflicts in objectives of stakeholders (Carton and Thissen, 2009). The above emphasizes the importance of tools to support negotiations among stakeholders that use maps in addressing the identification and resolution of conflicts about future land use. Participatory approaches to support face-to-face group decision-making comprise map-based tools, SDSS, information visualization, collaborative environments and the use of interactive instruments. Recent approaches include: interactive visual displays that link maps and MCA, e.g. CommonGIS (Andrienko and Andrienko, 2003); integrated GIS-MCA tools that focus on identifying conflicts and expressing priorities, e.g. the 'Spatial Group Choice' tool (Jankowski and Nyerges, 2001); Group-based SDSS (GSDSS) (Feick and Hall, 2002); tools to support negotiated land allocation, e.g. the Land-Use Planning Information System (LUPIS) (Recatalá Boix and Zinck, 2008) or Janssen et al. (2006); and visually-enabled 'geocollaboration' to support group work (MacEachren et al., 2005).

## 3. The approach

Our approach to structure a planning process involves face-to-face workshop meetings with stakeholders (public, private and research experts). It is based on a series of three interconnected participatory workshops: design, analysis and negotiation. The series starts with a design workshop, which involves the generation of reference alternatives, an inventory of the relevant objectives of all stakeholders and selection of evaluation criteria linked to these objectives. The design workshop is followed by an analysis workshop. Information about the region is presented in several ways to experts with different backgrounds to increase the understanding of the decision problem. The third of

the series is the negotiation workshop, in which stakeholders are invited to collectively improve a reference plan by negotiating land use changes. This workshop involves the value maps produced in the first two workshops and map-based MCA tools implemented in the Touch table to support the negotiations. The workshop setup comprises hardware and software. Hardware includes a laptop, the Touch table and a separate monitor screen (Figure 1). Software comprises MCA tools for dynamic plan evaluation, tools to support tradeoff identification and drawing tools to change land uses on the map. The tools were developed with CommunityViz Scenario 360 (http://www.communityviz.com/).

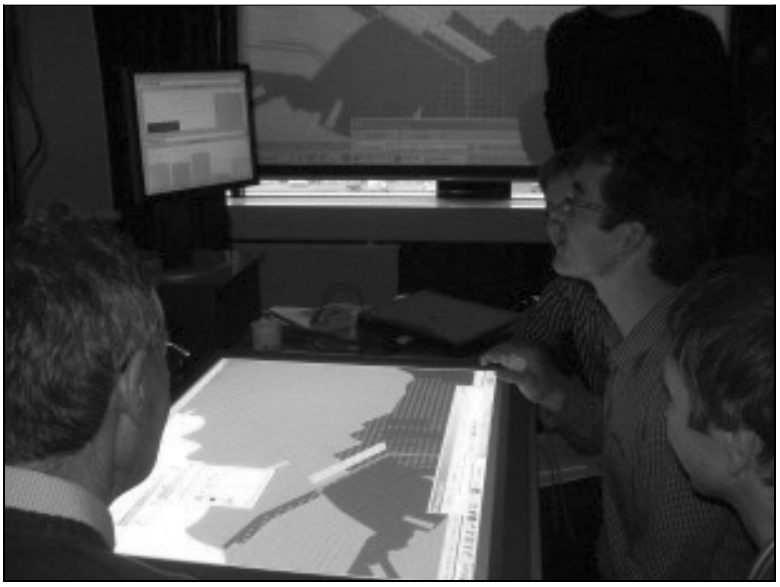


**Figure 1.** Hardware setup: Touch table and separate screen with support information

## 4. Implementation

We developed a negotiation exercise and informally tested it as part of the land use planning process of Bodegraven, a peat meadow area in the Netherlands with multiple stakeholders involved and a currently unsustainable land use situation that must be adjusted. Participants were experts of groups involved in research about peat meadows in the region. They were divided into two groups, each with a Touch table and asked to improve the same reference land use plan. Each group included three users, each playing a stakeholder role and associated with a land use type and evaluation objective (See Table 1). The purpose of the exercise was to compare both negotiation results. Both groups negotiated separately for the same length of time (one hour) and with the same goals. Records of the discussions reflected two different negotiation strategies, which resulted in two different plans with similar quality values.

**Table 1.** Tasks and goals of negotiation exercise

| Stakeholder | *Land use* | *Goal* |
|---|---|---|
| Nature organization | Nature | Achieve 860 ha of nature, trying to optimize the value for objective 'Nature' |
| Agricultural nature organization | Extensive agriculture | Achieve 1600 ha of extensive agriculture, trying to optimize the value for objective 'Landscape' |
| Farming organization | Intensive agriculture | With 3155 ha of intensive agriculture, land must be given in; try to maintain the value of 'Agriculture' as high as possible, keeping the best parcels and trading the worst ones |

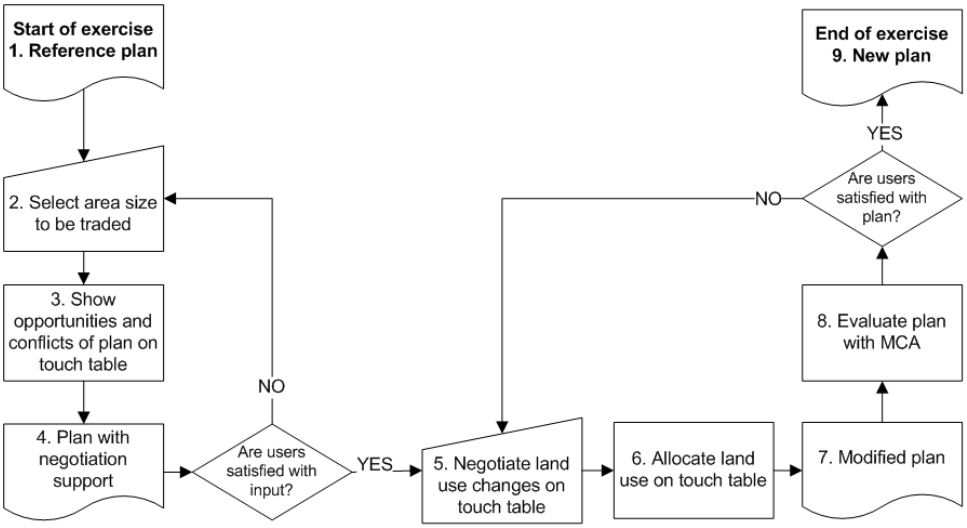



**Figure 2.** Flowchart of the process followed in a Negotiation workshop.

Figure 2 illustrates the flowchart of the workshop. A reference plan displayed on the Touch table is the starting point. The goal is to improve the plan, i.e. increase both its total and relative qualities, by changing land use of parcels and achieve a plan that is beneficial to all parties involved. Through a multi-user interface, the participants can retrieve parcels with either the highest or lowest value for the objectives concerned. The other users follow the same procedure so that opportunities and conflicts are simultaneously displayed on the map. This is done until sufficient 'negotiable' parcels are identified, i.e., parcels good for one stakeholder and bad for other (See Figure 3).

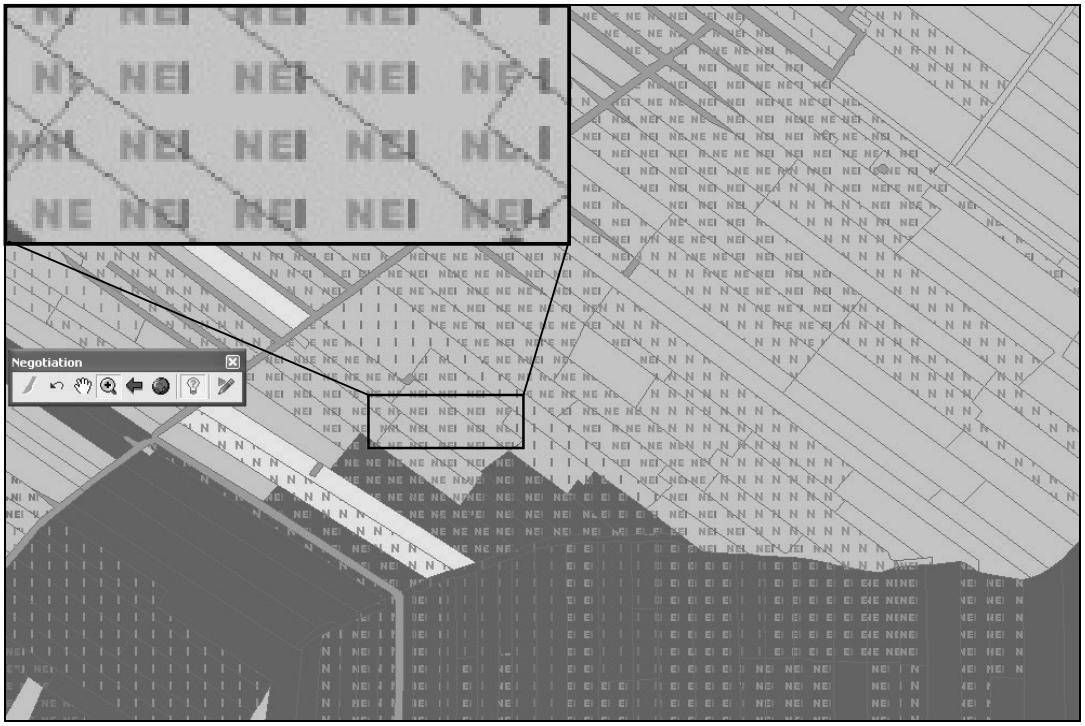



**Figure 3.** Land use map: nature (dark green), extensive agriculture (yellow), and intensive agriculture (light green) overlaid with tradeoffs as negotiation support: blue characters in zoomed-in inset show best areas for Nature and Extensive agriculture. Red characters show worst areas for Intensive agriculture.

Next, participants can focus on areas of interest and negotiate land use exchanges that are favorable. The next step is to implement these exchanges by painting new land uses on the map on the Touch table. With the drawing tools the participants can change land use of 'negotiable' parcels. Once the participants accept the trades and paint the changes, the MCA component reassesses the qualities of the modified plan (See Figure 4: 'Objectives Value'). Support information in the form of provisional evaluation results and the number of exchanged hectares is displayed on the separate screen as bar charts (Figure 4). The process of trading and reallocating continues until the participants are satisfied with the new plan and its qualities. This new consensus plan marks the end of the negotiation.

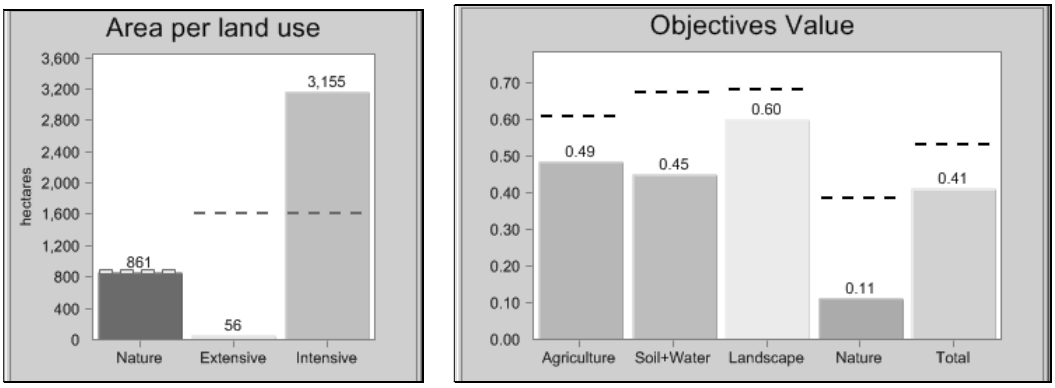



**Figure 4.** Support information: starting land use areas (right) and objective values for the negotiation exercise (left). Dashed lines on chart 'Area per land use' indicate target areas for each land use type. Dashed lines on chart 'Objectives value' indicate theoretical maximum values for each objective and total value of the plan.

## 5. Conclusions

This paper described an approach to support the negotiations involved in the design of a compromise plan. In a negotiation workshop, we provide participants with tools to disclose trade-offs between objectives of stakeholders as well as information about the overall and relative qualities of the plan that is being designed on the Touch table. The exercise setup allows optimizing individual objective values and the total plan. Interestingly, it was noticed that the participants used the support information provided on the table more than the summarized MCA information provided on the screen. Besides limitations such as number of users and land use types, in the evaluation module the value of one parcel is not influenced by neighboring parcels. It would be interesting to include aspects, such as homogeneity, connectivity, compactness, etc. into our evaluation framework. Further research is recommended on the incorporation of such spatial aspects into this type of participatory methods. Participants of the negotiation exercise were research experts involved. The next step is to test the negotiation exercise during a negotiation workshop with both private and public stakeholders and see whether they are still willing to negotiate within our framework with real interests at stake. This is expected to be more difficult because it will involve more direct stakes and narrower scopes for negotiation. While the negotiation support showed potential usefulness and the participants enjoyed the workshop, we will develop formal ways to test its effectiveness.

## References

Andrienko, N. and Andrienko, G. (2003). Informed spatial decisions through coordinated views. *Information Visualization*, **2 (4)** pp 270-285.

Carton, L. J. and Thissen, W. A. H. (2009). Emerging conflict in collaborative mapping: Towards a deeper understanding? *Journal of Environmental Management*, **90 (6)** pp 1991-2001.

Feick, R. D. and Hall, G. B. (2002). Balancing consensus and conflict with a GIS-based multi-participant, multi-criteria decision support tool. *GeoJournal*, **53 (4)** pp 391-406.

Jankowski, P. and Nyerges, T. (2001). GIS-Supported Collaborative Decision Making: Results of an Experiment. *Annals of the Association of American Geographers*, **91 (1)** pp 48-70.

Janssen, M., Goosen, H., and Omtzigt, N. (2006). A simple mediation and negotiation support tool for water management in the Netherlands. *Landscape and Urban Planning*, **78 (1-2)** pp 71-84.

MacEachren, A. M., Cai, G., Sharma, R., Rauschert, I., Brewer, I., Bolelli, L., Shaparenko, B., Fuhrmann, S., and Wang, H. (2005). Enabling collaborative geoinformation access and decision-making through a natural, multimodal interface. *International Journal of Geographical Information Science*, **19 (3)** pp 293-317.

Malczewski, J. (2006). GIS-based multicriteria decision analysis: a survey of the literature. *International Journal of Geographical Information Science*, **20 (7)** pp 703-726.

Recatalá Boix, L. and Zinck, J. (2008). Land-Use Planning in the Chaco Plain (Burruyacú, Argentina): Part 2: Generating a Consensus Plan to Mitigate Land-Use Conflicts and Minimize Land Degradation. *Environmental Management*, **42 (2)** pp 200-209.

## Biography

***Gustavo Arciniegas*** *holds B.Sc. and M.Sc. degrees in civil engineering and geoinformatics. His research interest is the use of spatial information technologies for land use planning, hazard monitoring and environmental problem solving. Currently he works as Ph.D. researcher in spatial decision support tools at the Institute for Environmental Studies, Amsterdam.*

***Ron Janssen*** *specializes in decision support for environmental management. His research topics are decision analysis and spatial analysis/evaluation, focusing on the effective use of information and the development of spatial decision support systems. Currently he is head of the department 'Spatial analysis and decision support' within the Institute for Environmental Studies, Amsterdam.*

# Identifying Dutch elm disease 'danger-spots' on the Isle of Man with an agent-based model

Bruce Mitchell [a], Joana Barros [b], Daniel Wendel [c]

Department of Geography, Environment and Development Studies,
Birkbeck, University of London,
Malet Street, London, WC1E 7HX

Telephone: (+440) 207 079 0644
Fax: (+44) 207 6316 498
Email a: bruce.birkbeck@ntlworld.com
b: j.barros@bbk.ac.uk c: djwendel@gmail.com

Keywords: Dutch elm disease control, agent-based modelling, local vulnerability

## 1. Introduction

The paper presents an agent-based model (ABM) of the spread of Dutch Elm Disease (DED), applied to the Isle of Man (IoM). The objective is to provide the Manx forestry authority with a tool to support their DED control campaign.

IoM has an estimated population of 250,000 elm trees, and is unusual in that DED – transmitted by the bark beetle Scolytus – is still being successfully fought. A strict control program was established soon after the disease arrived in 1992. So far, just over 1,000 trees have been lost, compared with around 30 million on the British mainland.

DED has been studied since 1918, and its epidemiology and lifecycle are by now well known. However, most DED studies are essentially aspatial. Models of DED have been developed, but focus on either the biological aspects of the disease – (Castro and Bolke (2004)) or on the spread of the disease (Swinton and Gilligan, 1996).

The present study proposes a three-dimensional spatial analytical agent-based-model (ABM) approach to identifying 'danger-spots' – locations on the IoM where outbreaks might lead to the greatest mortality among the Isle's elm population.

## 2. The DED Model

A prototype model was built in StarLogo TNG (MIT Media Laboratory, 2008), which supports modelling in three dimensions. This feature, together with the ease with which a model may be designed and implemented, makes StarLogo TNG particularly suitable for the project. Agent-based modelling was suitable for this project as it provided the opportunity to study the agents' behaviours independently as well as collectively.

The model emulates real-world processes of DED, with agents representing trees, foresters, beetles and birds (all of which influence the spread of the disease) embedded within a modelled 3D environment (**Figure 1**), based on a 75m digital elevation model (DEM) of the IoM's terrain and resampled to 400 metre raster cell size

When the model starts, agents Elm (800 units), Forester (2), Boid (4) and Beetle (variable) are generated and deployed across the landscape, their dispositions governed by random probability. Beetle agents then issue from infected elms (up to 1.5% of the 800) and seek out nearby healthy Elm in which to lay eggs and propagate the infection. Foresters seek to fell diseased Elm before the next generation of Beetle hatches. Boids pursue and eat Beetle on the wing. Where appropriate, these behaviours are modified by a 'sense' of where the respective prey is. The behaviour of each agent and their interactions is detailed in the flow chart (**Figure 2**).

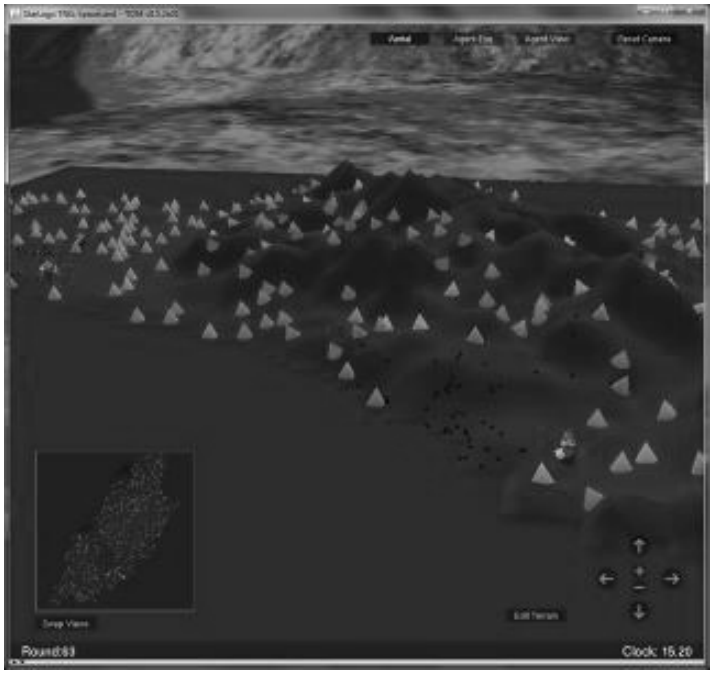



**Figure 1:** Agents Forester, Boid (blue), Beetles (black), and Elm.

The model proceeds for 1,000 game turns (GT), recording and exporting results as it goes. At GT60, it executes a census of the beetles – which always become established at a single location – determining their number, and the median centre and radius of their cluster. At GT 1,000, it resets and begins a new run. The model is run repeatedly to obtain a range of results and provide data for spatial analysis.

The modelled speed of DED spread and the motion of the beetles were informed by reference to Forestry experts and literature. The interactions of all agents were calibrated to minimise population explosions and collapses and to optimise performance and stability. Neither *gameturns* nor *runs* represent any particular unit of time. Each run therefore simulates no more than one possible outworking of interactions between landscape dynamics and multiple agents.

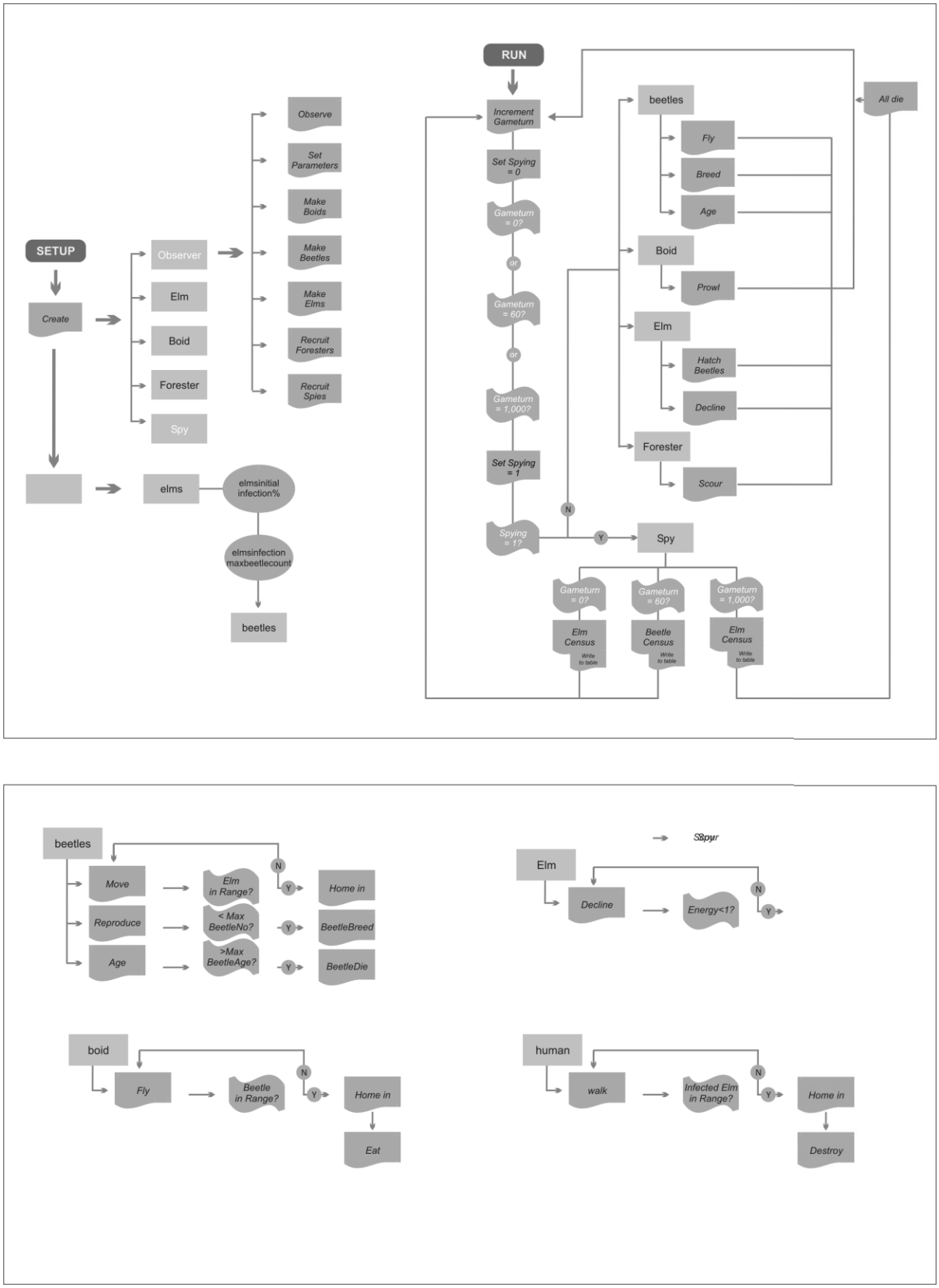



**Figure 2:** DED Model flowchart

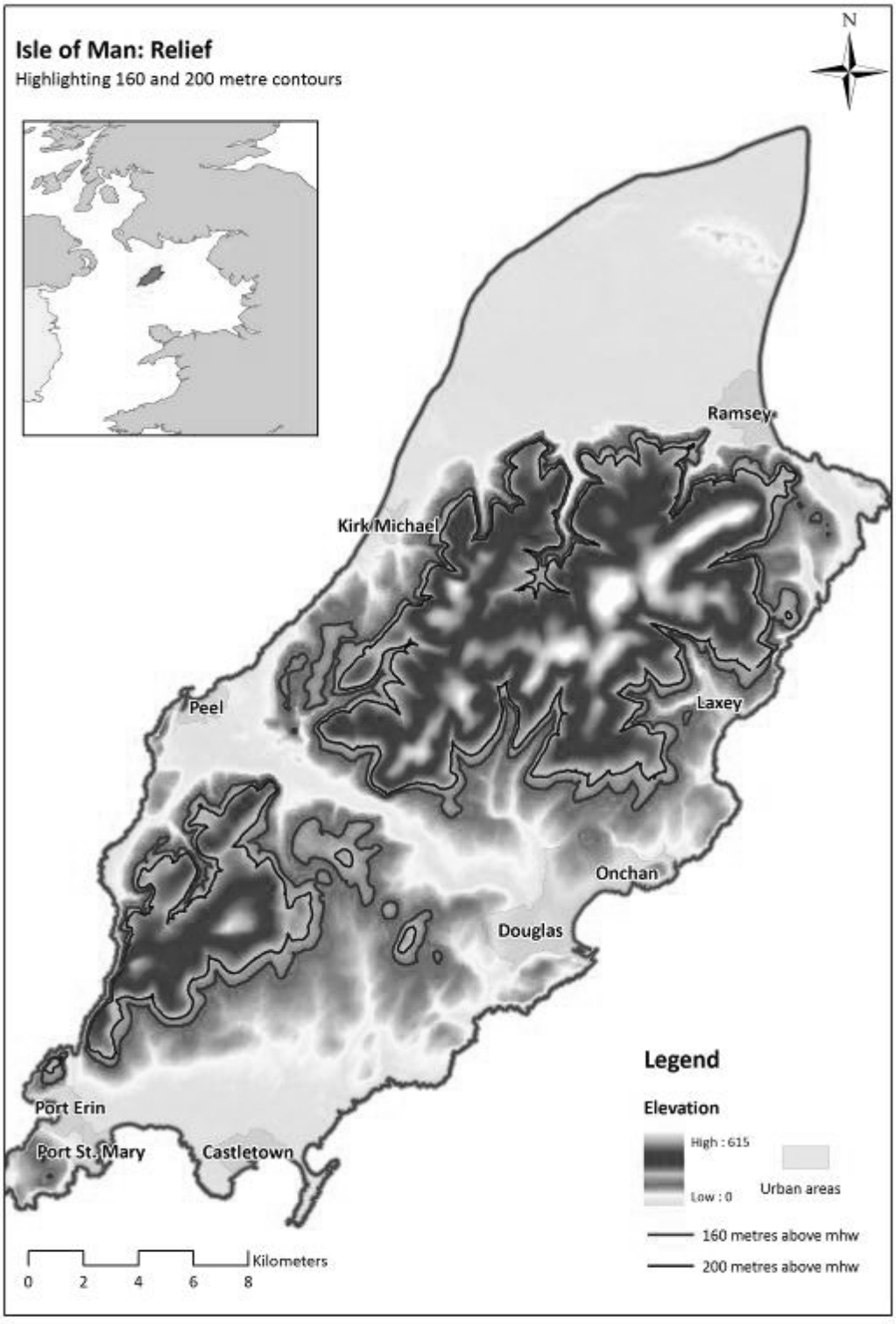



**Figure 3:** Isle of Man: location and relief

## 3. The impact of the landscape on DED dynamics

The principal idea explored by this study is that, dependent upon terrain, infestations originating in one part of the island might be contained, or find easy passage to populations of elm beyond. It is proposed that a correlation may exist between original epicentres of infection, terrain and the number of elms destroyed. This is based on the premise that Scolytus beetles do not fly much beyond the maximum elevation locally colonised by elms (the 'elm-line') - on the grounds that they are unlikely to stray far from their food source – and that landscape above the elm line (which will vary from place to place) can therefore act as a *cordon sanitaire* for DED control strategies. This concept lies behind the East Sussex Control Zone, a DED control program which has, since the early 1970s, exploited higher ground to the north of Brighton and Hove.[1] Unfortunately, enquiries with Forest Research failed to turn up any data on beetle flight habits and suggested that no such data exists. Nevertheless, the success of the East Sussex Control Zone, and the historical pattern of spread on the IoM, offer persuasive support.

On the Isle of Man, two ranges of hills (**Figure 3**) exceed the local elm-line – which rarely exceeds 150 metres (Fairhurst (1982)) – and restrict the flow of the disease across the island. The elm-line has been fixed in the model at 160 metres. A ceiling on beetle flight (beetle-line) has been set at an arbitrary 25 % higher (200m).

The combination of DEM, elm-line and beetle-line, together with parameters *beetlemaxrange* and *beetlemaxage* places elevation at the model's centre. The terrain and initial disposition of elm will result in beetles operating out of any disease epicentre finding some areas accessible, others out of range.

## 4. Modelling vulnerability

Each cell in the DEM was accorded a Raster Cell Neighbourhood Statistic (RCNS) to reflect its degree of openness to being traversed by Agent Beetle and, consequently, the level of risk of any local Agent Elm. Every cell representing land at or below the beetle-line was given the value 1 (i.e. may be traversed by agent beetle). All higher cells were given value 0. A 5x5 cell grid was then passed over the DEM and the central cell received the sum of the 25 cells. A second passage of the 5x5 grid produced the median to these sums – the **RCNS**. Adjacent cells with the same RCNS form **zones.**

When the RCNS is classified in eight natural breaks (**Figure 4**), the most vulnerable class of zones falls into five **regions**. Disease may potentially spread widely *within* the two largest regions, but transit *between* these two regions will be more difficult, as the beetles will have to pass through zones with lower RCNS values (i.e. with fewer passable cells). The disease may therefore remain trapped within these regions.

[1] See *http://www.eastsussex.gov.uk/environment/woodlands/dutchelms/default.htm*

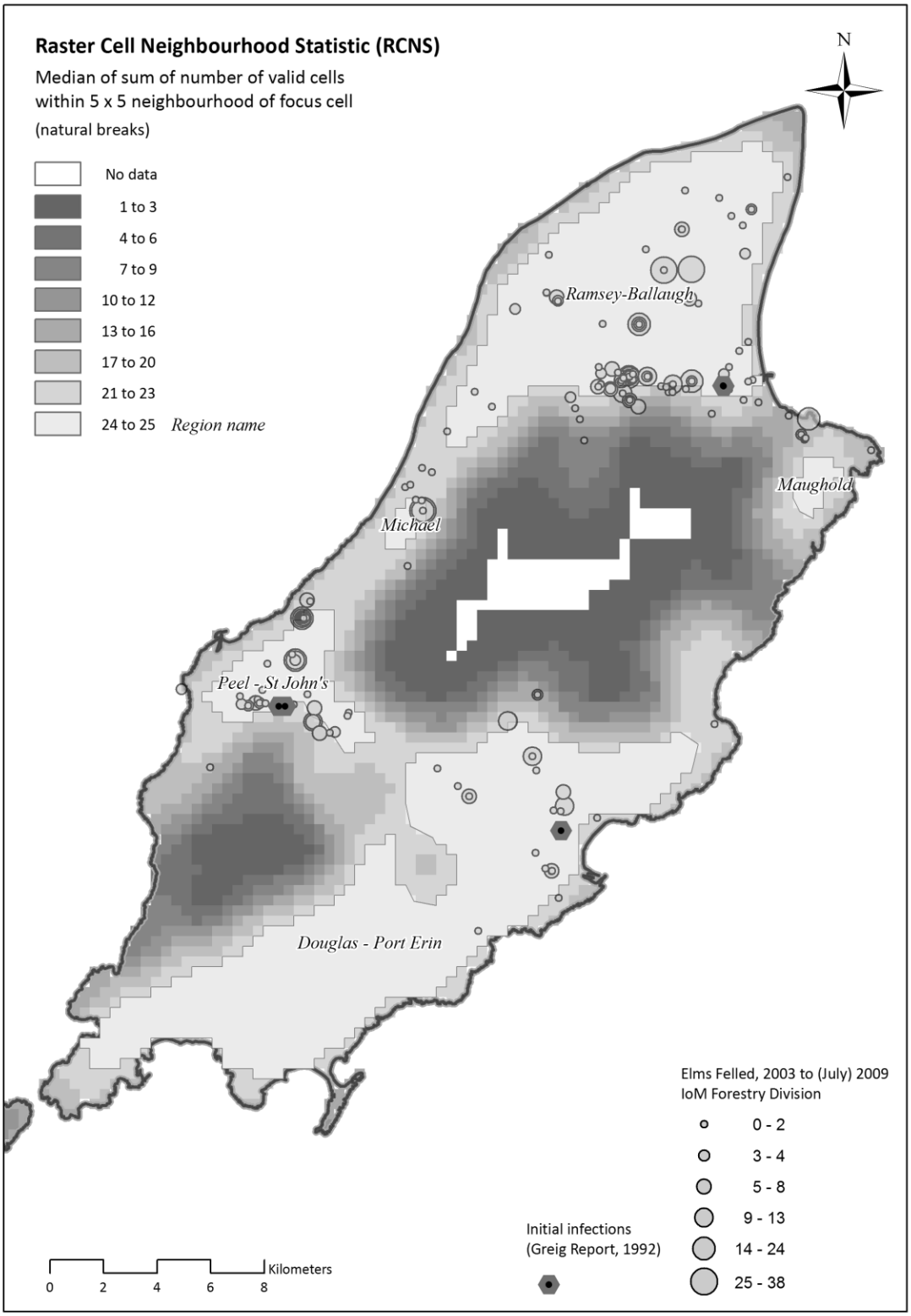



**Figure 4**: RCNS Regions

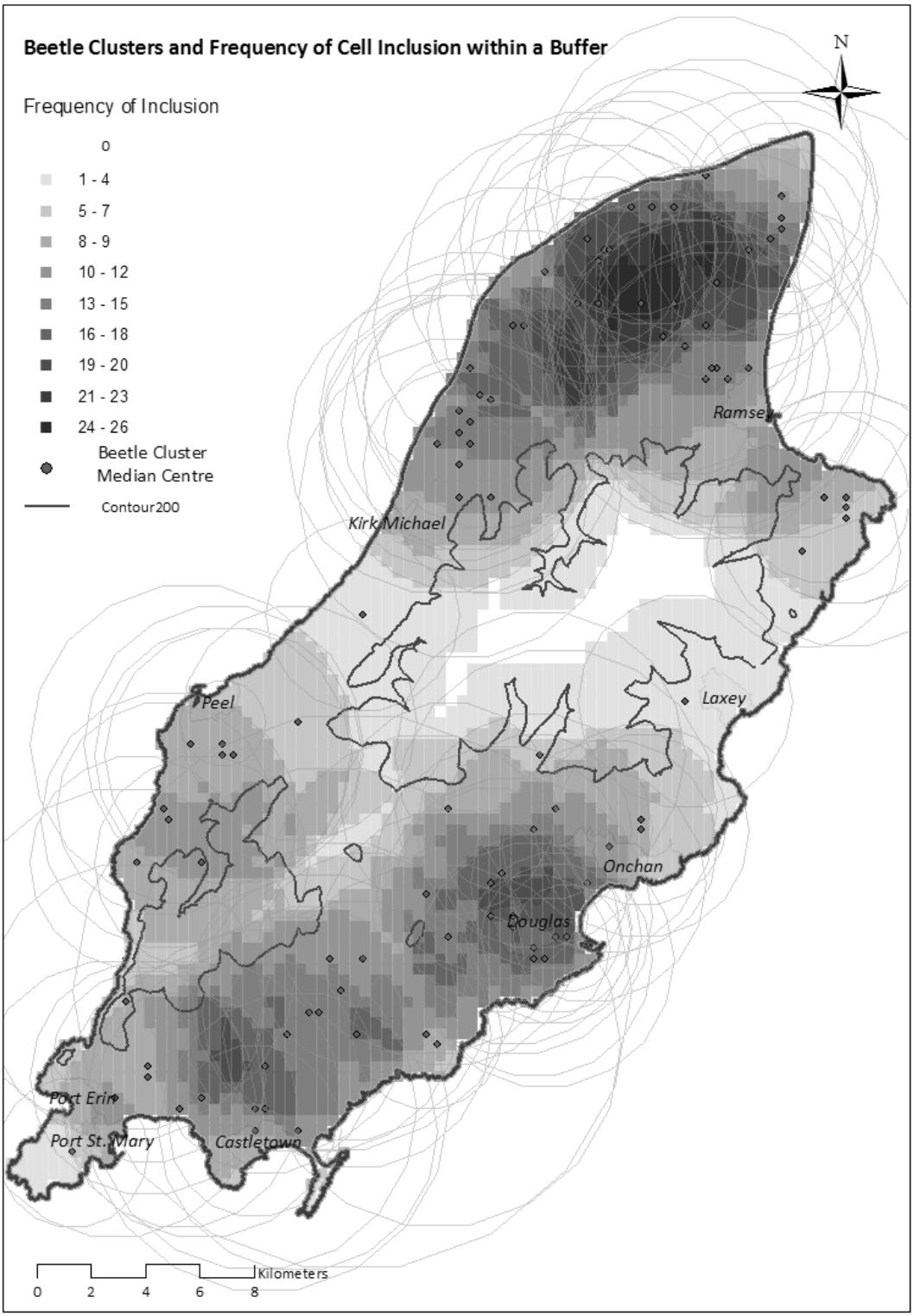



**Figure 5:** Beetle Clusters and Frequency of Cell Inclusion within a Buffer

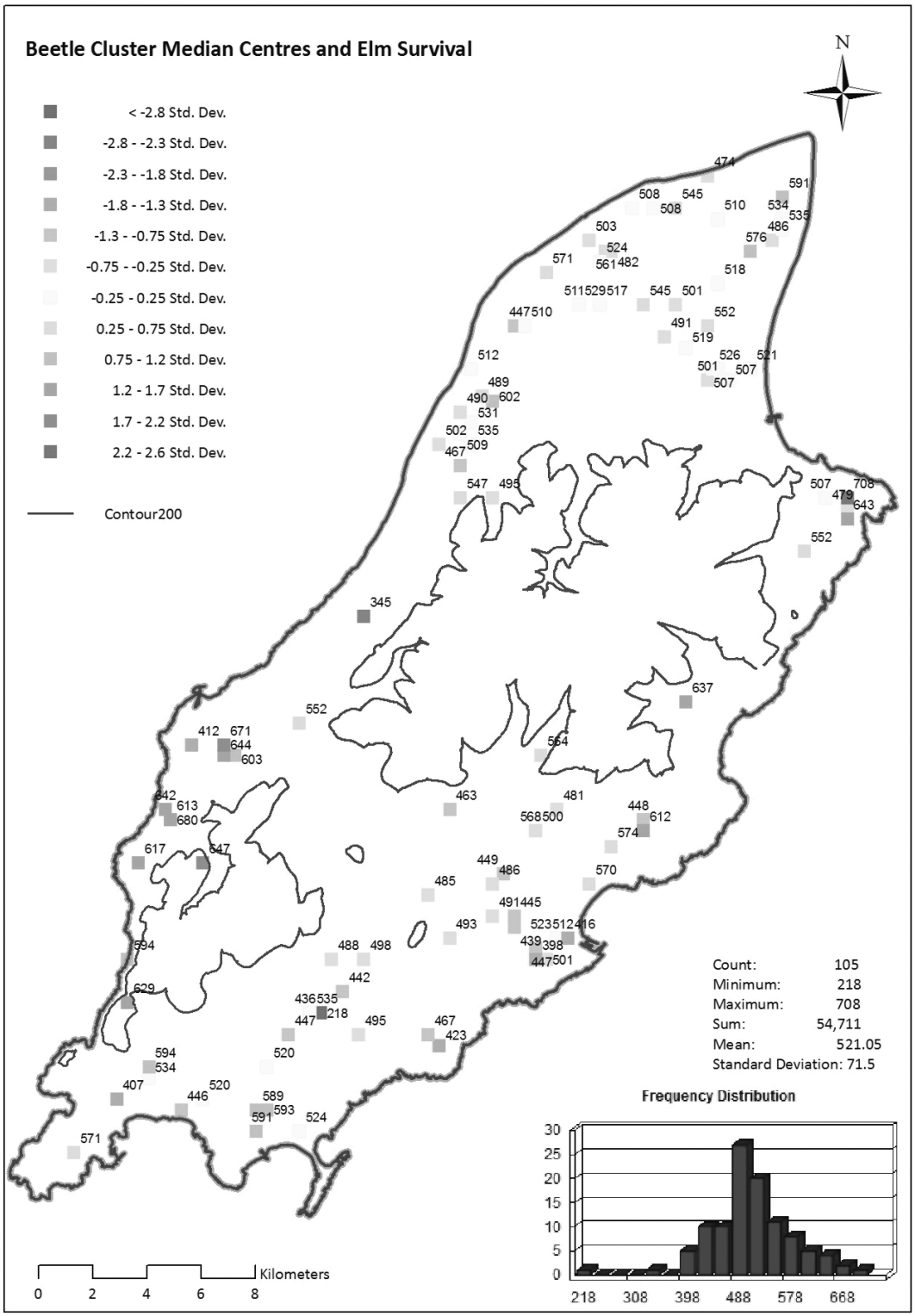



**Figure 6**: Beetle Cluster Median Centres and Elm Survival (Gameturn 1,000)

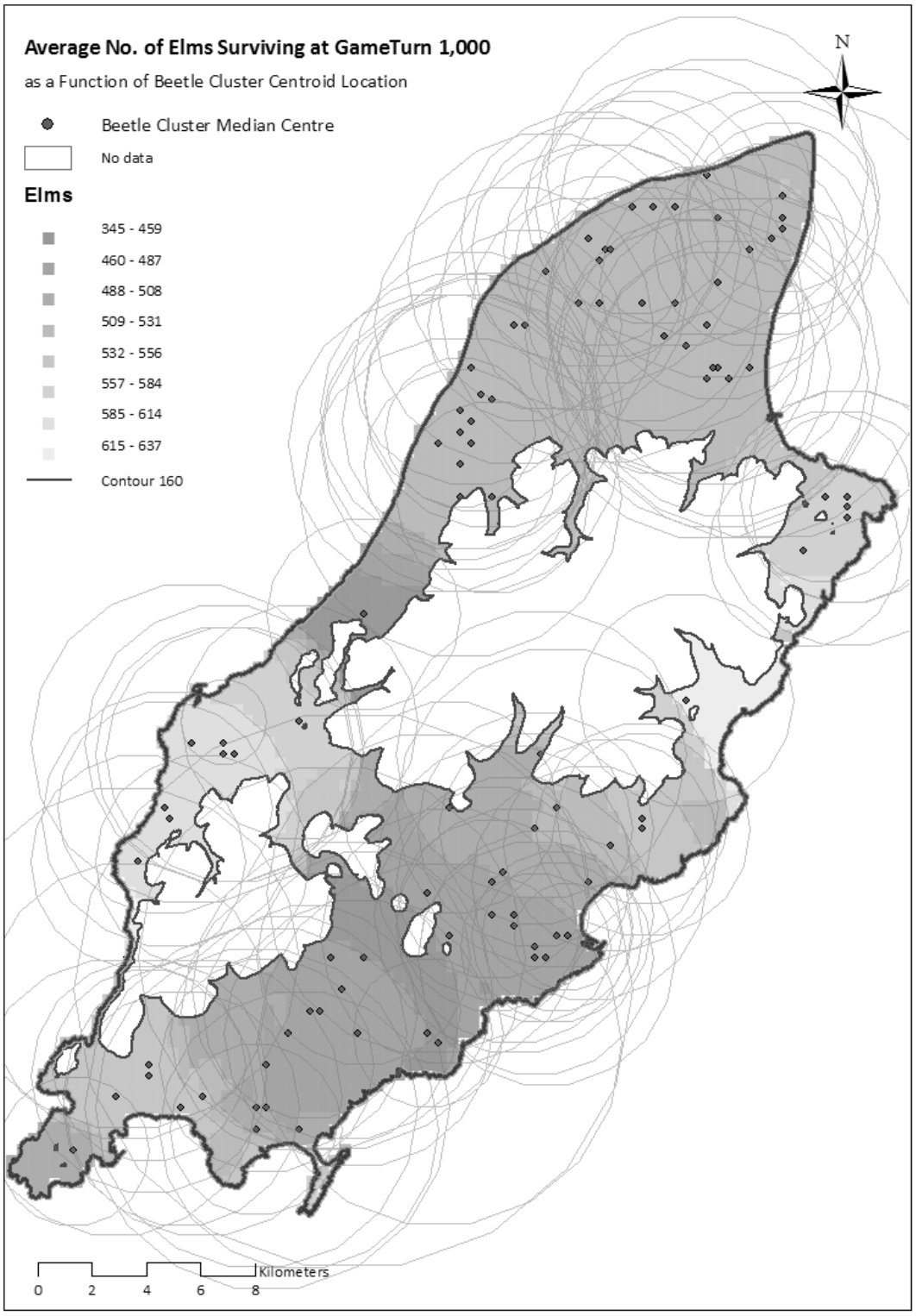



**Figure 7:** Danger-spots - Elms surviving to GT1000 (mean) as function of cluster location

## 5. Discussion of results

Each run completed by the TNG model represented an outbreak of the disease originating in a random location. Each generated a different (GT60) beetle cluster buffer and final (GT1000) outcome. In all, 106 runs were successfully completed. Often, the beetles failed to establish themselves and the run terminated prematurely. The great majority of successful runs began in zones with the highest RCNS. Results obtained from the hillier districts are dependent upon fewer results due to the small number of runs completed.

**Figure 5** shows the frequency with which individual cells appear in GT60 buffers. Three areas stand out: near Douglas, north of Castletown and, most prominently, in the northern plain.

In **Figure 6** (standard deviations), we see that the location of the beetle cluster has a marked effect on the number of elms that survive. The majority of clusters in the northern plain failed to break through the pinch points even to the elm populations of Laxey and Peel, explaining a generally moderate mortality of less than one SD from the mean. Clusters near Douglas, with easier access to the secondary region of Peel, were generally more destructive: those located near the hills, less so. **Figure 7** translates this into the number of elms that on average survived when any given cell was included in a cluster. The darker shades signify areas which produce the highest mortality, and are therefore the danger-spots we have been looking for.

## 6. Conclusions

Preliminary results conform to the expected pattern – that parts of the island (in the presence of a randomly distributed elm population and the absence of local climatic variations) may be more conducive to beetle generation and the onward spread of Dutch elm disease than others.

Next steps include:

- Increase in the number of runs to enhance stability of results;
- Use of real data on the spatial distribution of the elm population;
- Elimination of all non-essential variability within the model; and,
- Model upgrade to a dedicated ABM framework e.g. REPAST.

## 7. Acknowledgements

**Office for National Statistics**

Funding body

**Birkbeck College, University of London**:

Dejan Markovic, Dennis Rotheray, and Simon Measho, co-authors of the initial version of this model

**Isle of Man**

Forestry Division, Dept of Agriculture, Fisheries and Forestry, St John's.

IoM Survey Mapping Service, Department of Local Government & the Environment, Douglas.

**Forest Research**

Dr David Rose.

## 8. References

Brasier, C. M. (1996). *New horizons in Dutch elm disease control*. In *Report on Forest Research, 20–28.* HMSO, London.

de Castro F and Bolker B (2005). *Mechanisms of disease-induced extinction*. Ecology Letters, (2005) 8: pp 117-126

Fairhurst, C. (unpublished 1982). *The Elms of Man.*

Grieg, BJW (unpublished, 1992). *Dutch Elm Disease - Isle of Man*. Forest Research, Disease Diagnostic and Advisory Service, PAT 287.

Mitchell, B. et al:. *Dutch elm disease on the Isle of Man – Identifying 'danger-spots' using an agent-based model*. GeoComputation 2009, Sydney, Australia. Available online at: *http://www.biodiverse.unsw.edu.au/geocomputation/proceedings/PDF/Mitchell_et_al.pdf*

MIT Media Laboratory (2008-9). *StarLogo TNG 1.2* (Software). Available online at *http://education.mit.edu/starlogo/*

Rohani P, Keeling MJ (2002). *Estimating spatial coupling in epidemiological systems: a mechanistic approach*. *Ecology Letters*, 5: pp 20-29

Rose, D. 2000. *Dutch Elm Disease in the Isle of Man. A review of current practices and recommendations for the future*. Forest Research, Disease Diagnostic and Advisory Service.

Swinton J, Gilligan CA (1996). *Dutch Elm Disease and the Future of the Elm in the UK: A Quantitative Analysis.* Philosophical Transactions, Biological Sciences. Vol. 351, No. 1340. The Royal Society.

University of Chicago (2009). *RePast Simphony 2.0* (Software). Open source. Available online at *http://sourceforge.net/projects/repast/*

## 9. Biographies

*Bruce Mitchell has just completed a distance-learning MSc in GISc at Birkbeck, University of London. He is employed in the Data Visualisation Centre, Methodology Directorate at the Office for National Statistics. His research interests extend beyond GISc to geography, history, languages and forestry.*

*Joana Barros is a lecturer in GI Science and the MSc in GI Science Programme Director at Birkbeck, University of London. Her research interests include computational models of geographical systems, agent-based simulation models of urban systems, as well as urban growth and change in developing countries.*

*Daniel Wendel (MIT) has led the StarLogo TNG development team at the Scheller Teacher Education Program since 2007. As a member of STEP, Daniel's research interests centre around educational technologies, especially games and simulations.*

# vizLegends : Re-Imagining Map Legends with Visualization

Jackie Clark[2], Jason Dykes[1*], Fiona Hemsley-Flint[2], David Medyckyj-Scott[2], Lasma Sietinsone[2], Aidan Slingsby[1], Tim Urwin[2], Jo Wood[1]

[*] *corresponding author*

[1]giCentre, Dept. of Information Science, City University London, EC1V 0HB
Tel. +44 (0)20 7040 8906
jad7 | a.slingsby | jwo @soi.city.ac.uk

[2]EDINA National Data Centre, University of Edinburgh, 160 Causewayside, Edinburgh, EH9 1PR
Tel. +44 (0)131 650 3302
jackie.clark | fiona.flint | l.sietinsone | t.urwin @ed.ac.uk; medyckyj-scottd@landcareresearch.co.nz

KEYWORDS: legend, cartography, visualization, design, Digimap service, online web mapping

## 1. Introduction

UK tertiary education accesses a large and diverse collection of spatial data through EDINA's Digimap service (Sutton et al., 2007). Digimap clients combine and present data of varying theme, content, scale and format in single maps. Legends are essential in this context but bulky and difficult to navigate if comprehensive. Accordingly we conduct a structured and applied study to explore possibilities for using visualization methods to revisit the map legend. Innovative candidate '*vizLegend*' designs are developed by re-imagining the legend in a collaborative three-phase process involving requirements, review, rapid prototyping, evaluation and redesign.

***Phase 1 – 'Imagination Exercise':*** to establish context, exchange ideas and data and develop broad requirements, guidelines from existing practice, themes that inform design and '*digital wireframes*' to demonstrate possibilities.

***Phase 2 – 'Constrained Development':*** to evaluate themes and their implementation in digital wireframes, develop more specific requirements and generate focussed '*digital prototypes*'.

***Phase 3 – 'Evaluation and Deployment':*** to evaluate prototypes, establish user responses and incorporate themes and functionality into Digimap services.

## 2. Imagination Exercise

### 2.1 Legend Requirements

Requirements were established at a structured one-day workshop. Current and prototype Digimap services were presented and novel InfoVis techniques introduced to stimulate creative thinking. Formal data collection followed, focussing on three characteristics of legends:

i. successes / problems in existing Digimap legends
ii. aspirational legend characteristics
iii. Digimap legends tasks and functionality

Responses from individual participants were aggregated, ranked and subsequently summarised in graphical tables. Figures 1 & 2 show examples for the final two characteristics. Indicative statements describe each response (or response set). These are ordered from top to bottom by 'strength of feeling' established at the workshop. In Figure 1 the numbers show progression through a four-stage hierarchical prioritisation process. In Figure 2 the numbers are rankings as determined through discussion within two groups of participants. Bar length is indicative of the number of times a broad characteristic was expressed individually at the outset in both cases.

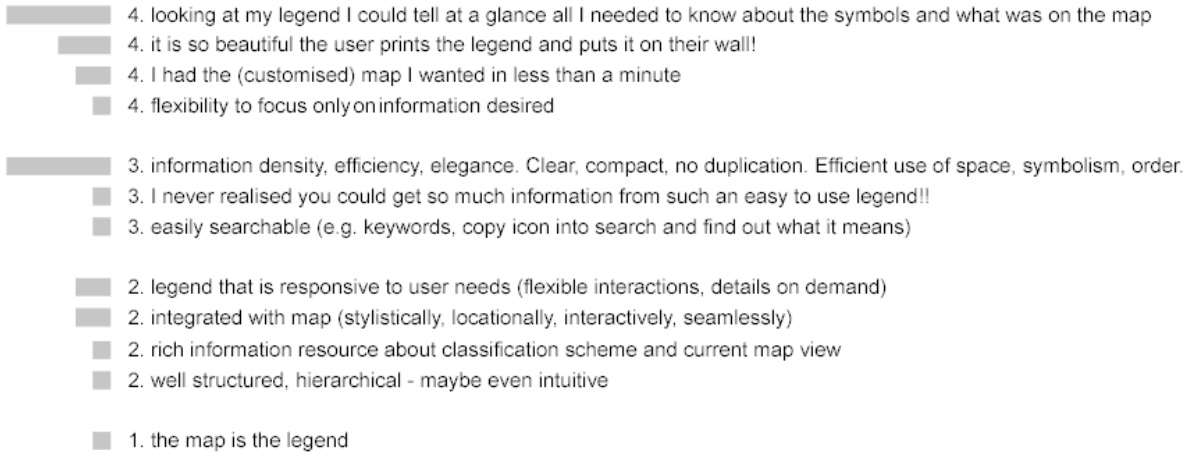



**Figure 1.** Aspirational Legend Characteristics.

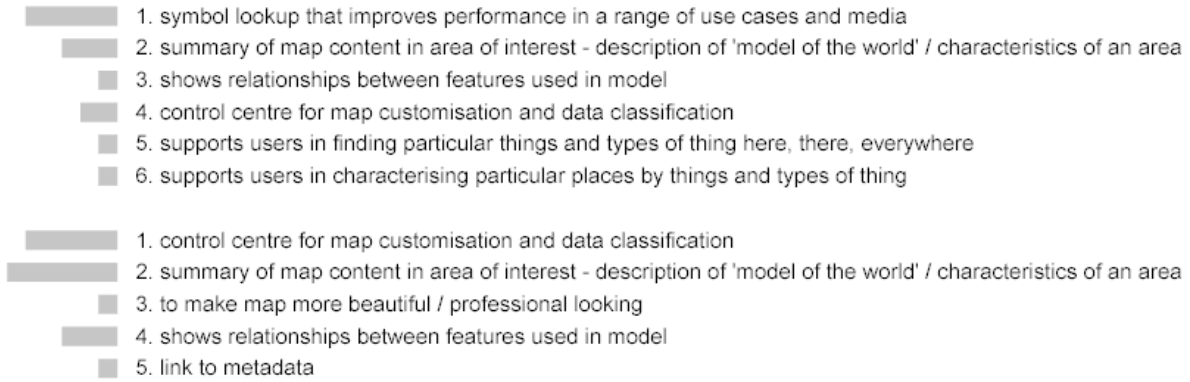



**Figure 2.** Tasks and Functionality – ordered response sets from two groups (top and bottom).

### 2.2 Legend Design

Legend design was subsequently informed by a review of 47 sources including cartography texts, journal papers and digital resources. Cartographic text books describe legends as fundamental map components that should include "critical" (Robinson et al., 1995) or "unknown" (Dent, 1990) information that is "not self explanatory" (Slocum et al., 2009). Experimental work shows positive effects of design on performance (Cox, 1976; DeLucia and Hiller, 1982; Aspaas and Lavin, 1989). Guidance on design is very general however, with occasional examples (Slocum et al., 2009) and alternatives (Cuff and Matson, 1982), but few principles, perhaps because it is difficult to generalize between experimental contexts. Six principles for vizLegends were derived from our review:

1. **Process** design the legend in the manner of a map. The seven controls on map design (Robinson et al., 1995) may be helpful.
2. **Selection** should not be comprehensive. Include items deemed 'critical', 'unknown' or 'not self-explanatory' to minimise map-legend references.
3. **Symbols** should relate directly to those mapped, each other (in terms of layout – see below) and the referent.
4. **Layout** should represent information structure (see 'Symbols'). Spatialization may be beneficial.
5. **Position** legends should be encountered before maps – this may vary with task, user, time and data set.
6. **Dynamism** should facilitate map-legend references and variation of selection, layout, symbolism and position as required (by user, data, task, etc.).

Concepts and approaches documented in the literature that may help achieve these principles include 'The Active Legend' (Sieber et al., 2005), bi-directional highlighting (Sieber et al., 2005) and the application of styles (Jolivet et al., 2008, 2009; CloudMade, 2009). Creative thinking, continuous redesign and "vigorous editing" (Brewer, 2005) are deemed to be essential in their application.

### 2.3 vizLegend Themes

These legend requirements and design principles provided impetus for a creative exercise in which a series of broad themes were developed to frame vizLegend ideas. Each theme describes a novel perspective on the legend that may address some of the established requirements through the identified principles. Themes are generic and may be used individually or in combination to guide the development of wireframes or prototypes. They are neither comprehensive nor mutually exclusive.

1. **The Map is the Legend**
   *The roles of map (spatial representation of geographic setting) and legend (spatial representation of map content and symbolism) are blurred.*
2. **The Legend as Statistical Graphic**
   *A graphical statistical summary of current map content and a query filter for map exploration.*
3. **The Scale Independent Legend**
   *Data integrated seamlessly from multiple sources with conflicts resolved.*
   *Sources identifiable but grouped by other characteristics.*
4. **The Relevant Legend**
   *Shows only what is required… as this changes. May apply to various other themes.*
5. **A Legend of Legends**
   *Alternative styles displayed with layout reflecting their relationships. Widely applicable.*
6. **Map of the Pops**
   *Legend items selected and arranged according user community needs. Widely applicable.*
7. **The Referent is 'Ground Truth'**
   *Symbols augmented or replaced with (local, community contributed) imagery.*
   *User community determines relevance. May be combined with various other themes.*
8. **My Legend, My Map**
   *User controls and saves content, layout and order according to task, knowledge, location, symbolism. Can be considered a subset of theme 4.*

### 2.4 vizLegend Digital Wireframes

The themes were used to develop four digital wireframes through which means of addressing the requirements were explored: *The Map is the Legend*; *The Legend as Statistical Graphic – Bar Chart*, *Matrix Plot*, *Hierarchy*. Each is named according to the dominant theme, but may be influenced by others. Requirements addressed by each wireframe were tabulated: the aspirational characteristics that '*The Map is the Legend*' wireframe was designed to accommodate are emphasized in Figure 3.

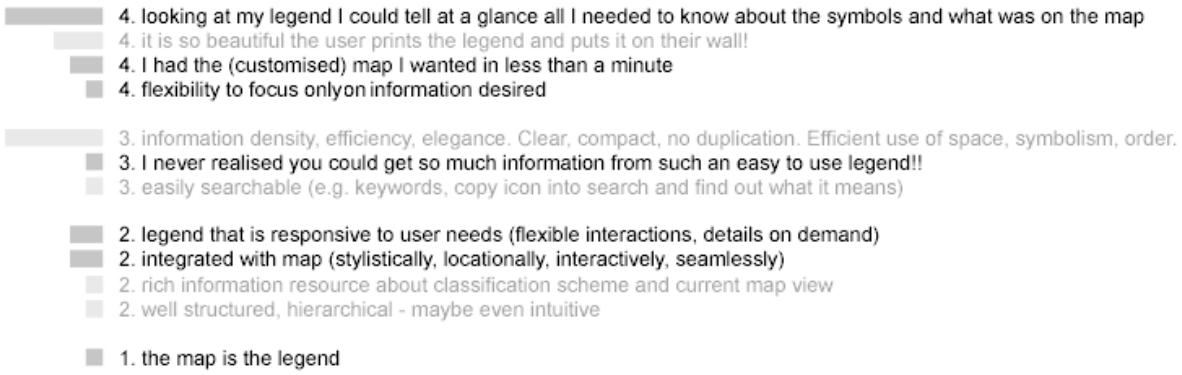



**Figure 3.** Aspirational Legend Characteristics: '*The Map is the Legend*'.
The requirement sets addressed in the wireframe are highlighted in dark grey.

The '*The Map is the Legend*' digital wireframe (Figure 4) transitions smoothly between alternative combinations of layout and selection with characteristics of themes 1, 4, 5 and 8:

- *Legend* - 1D ordered layout with single case of each feature (Figure 4, left)
- *Map* - 2D spatial layout of all features (centre)
- *Mapped Legend* - 2D spatial layout with single case of each feature at indicative location (right)

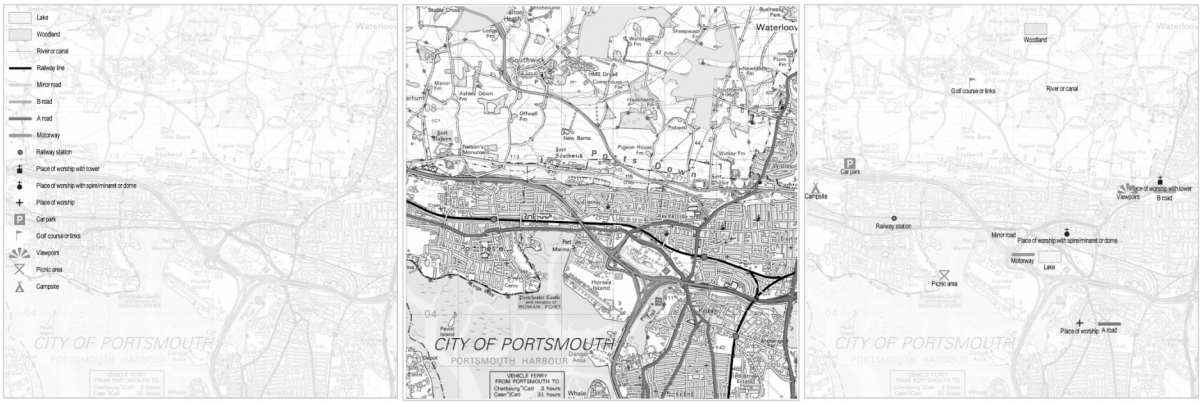



**Figure 4.** *The Map is the Legend*. Three states: legend, map, mapped legend (*left – right*).
*© Crown Copyright/database right 2009. An Ordnance Survey/EDINA supplied service.*

The '*The Legend as Statistical Graphic - Hierarchy*' (Figure 5) uses themes 1 and 2 to show hierarchical bedrock classification through a two-dimensional spatially ordered space-filling treemap (Wood and Dykes, 2008). Bi-directional interactions (Sieber et al., 2005) feature strongly in the wireframe and include zoom / pan to select spatially with legend updated according to map content.

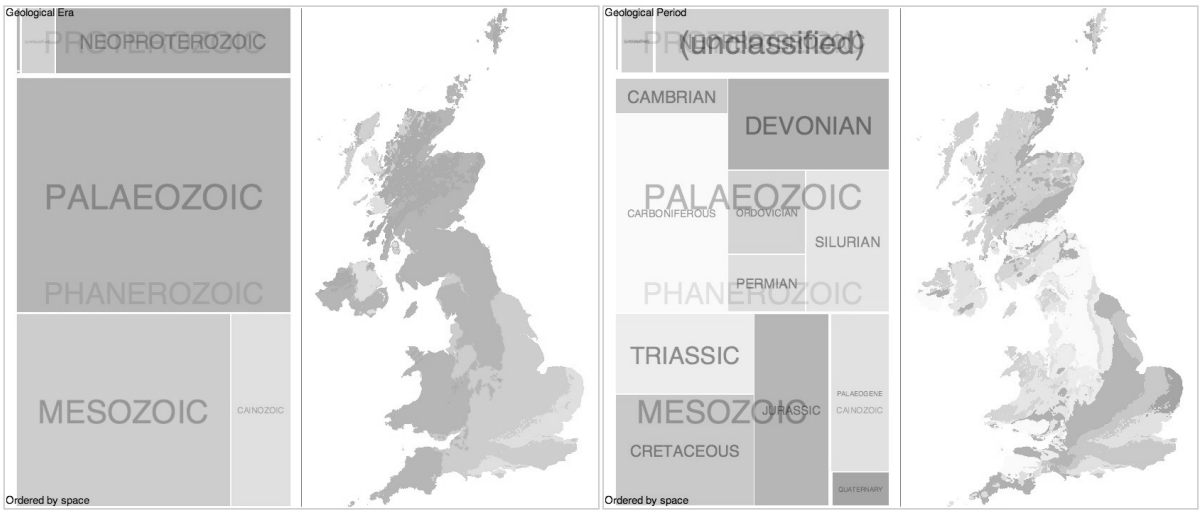



**Figure 5.** '*The Legend as Statistical Graphic – Hierarchy'*. Geological map in which areas on the legend relate to areas on the map for era (*left*) and period (*right*). Legend can be interactively reordered to show attribute, spatial or chronological orders at each level of the attribute hierarchy
*Geological Map Data © NERC 2009.*

## 3. Feedback and Digital Prototypes

The results of the 'imagination exercise' generated ideas for enhancements, developments and evaluation. These included an EDINA mock-up of theme 7 with community contributed styles and locally relevant photographs (Figure 6).

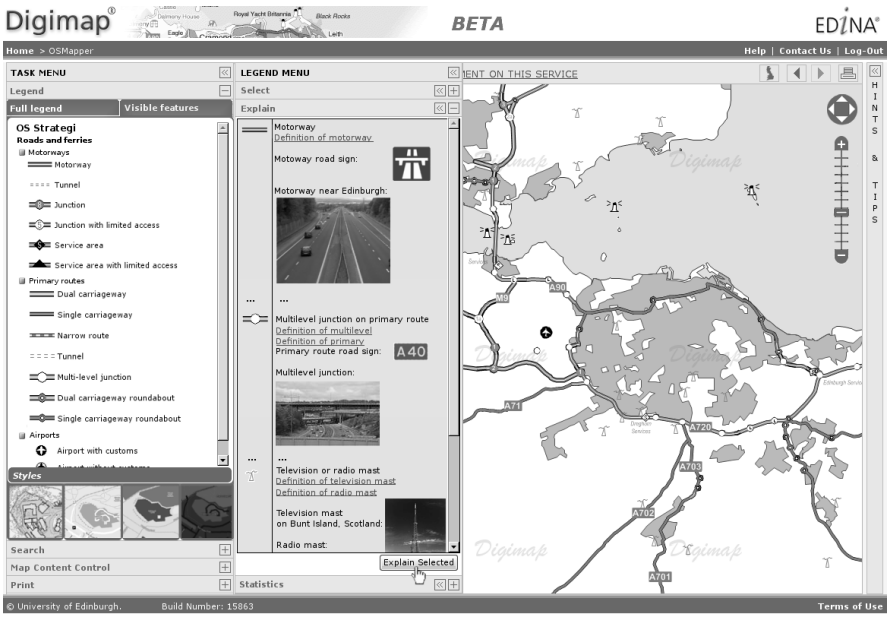



**Figure 6.** '*The Referent is Ground Truth*'. EDINA mock-up.
*© Crown Copyright/database right 2009. An Ordnance Survey/EDINA supplied service.*

Additional requirements were also derived following interaction with the digital wireframes: using *Strategi;* visualizing the *Strategi* feature hierarchy; extending bidirectional interactions. Two of the digital wireframes were enhanced accordingly. The resultant prototypes are shown in Figures 7 and 8.

## 4. Conclusion and Ongoing Work

Harrower (2003) contends that "*Digital and Web cartography fails when we try to reproduce paper maps on-screen*" and calls for a creative approach. We show that a structured collaborative design process with rapid iterative development that draws upon real data can result in innovative cartography and candidate solutions that have potential. Our re-imagination exercise identifies various new roles and approaches for legends that may be suitable for Digimap and elsewhere through usable principles, themes and prototypes. These are implementable and can be evaluated.

Visualization methods are best communicated through interactive media and evaluated by map users. The vizLegend themes and prototypes will be further investigated by the Digimap user-base and subsequently considered for use in Digimap clients as we evaluate prototypes, establish user responses and consider incorporating themes and functionality into Digimap services.

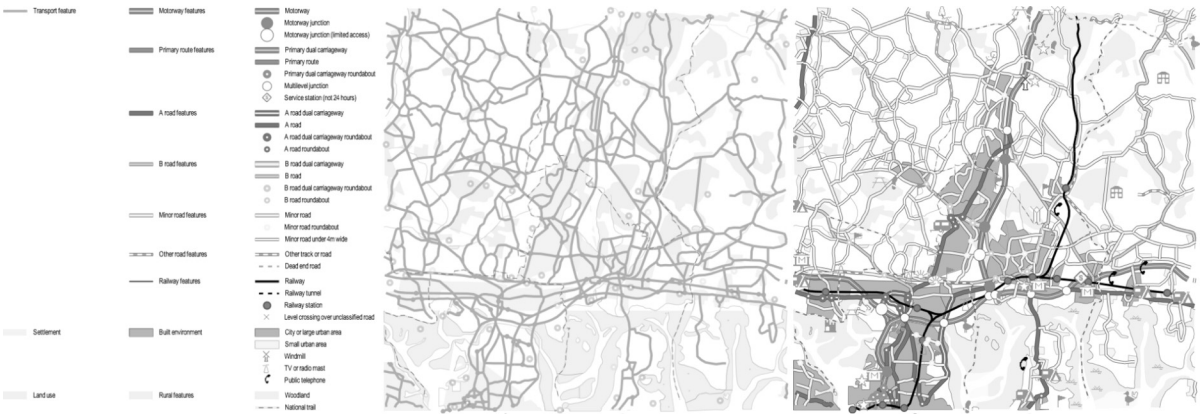


**Figure 7.** '*The Map is the Legend*'. Digital prototype with legend items arranged hierarchically (left), styled hierarchically (centre) and styled according to current Digimap convention (right).
*© Crown Copyright/database right 2009. An Ordnance Survey/EDINA supplied service.*

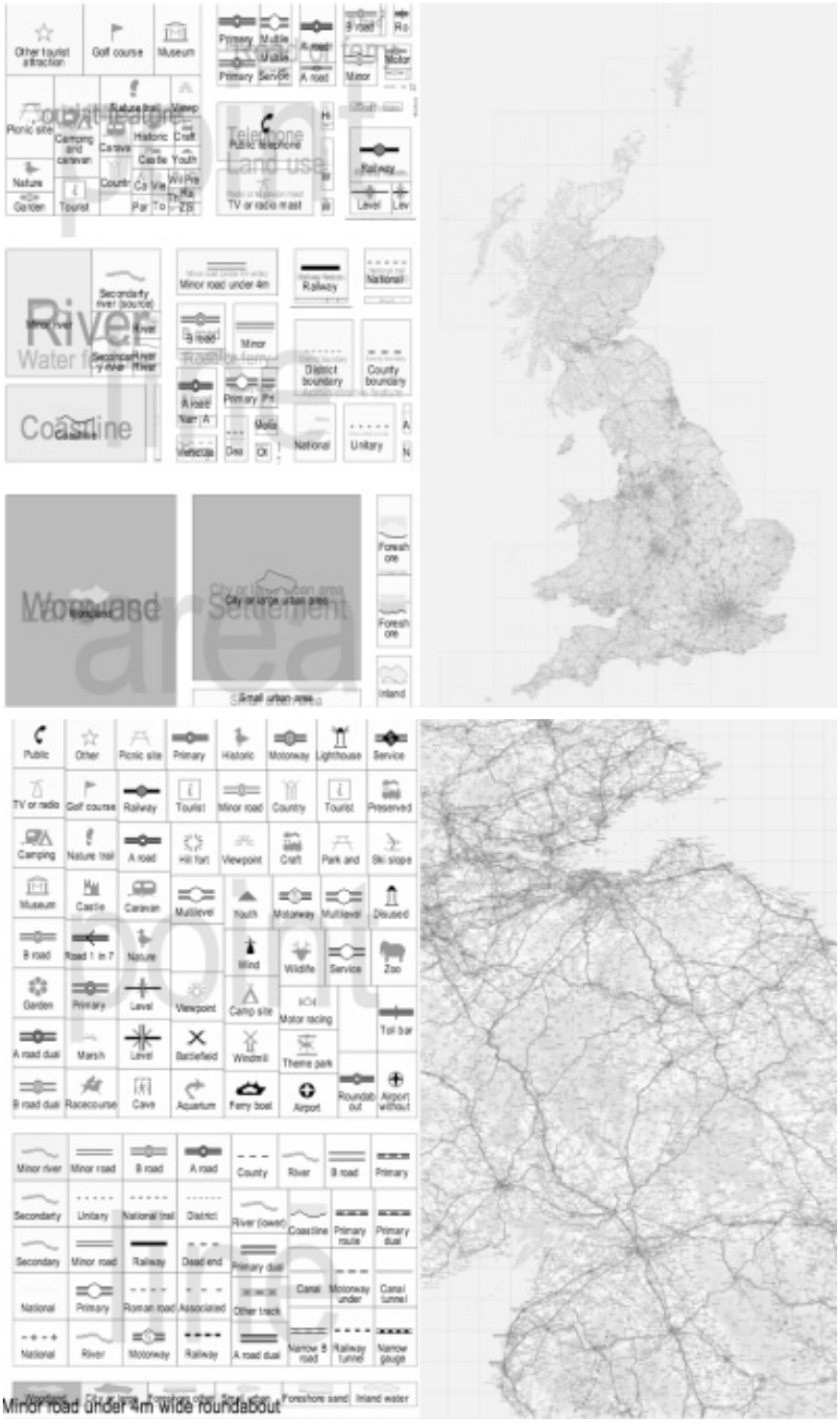



**Figure 8.** '*The Legend as Statistical Graphic - Hierarchy*'. Digital prototype showing feature hierarchy and occurrences of point, line and area features in national Strategi data set using area and colour (top) and colour (bottom) at national (top) and interactively determined local (bottom) level. Raster backdrops are generalized in the prototype to ensure rapid response.


## 5. Acknowledgements

EDINA National Data Centre funded this work through its small-scale development budget. Data provided by Ordnance Survey GB and the British Geological Survey. Thanks to EDINA Geodata Services and User Support for contributing to the three vizLegends workshops.

## References


Aspaas, H.R. & Lavin, S.J. (1989). *Legend Designs for Unclassed, Bivariate, Choropleth Maps*. CAGIS, 16(4), 257-268(12).

Brewer, C.A. (2005). *Designing Better Maps; A Guide for GIS Users*, Redlands, CA: ESRI Press, 205pp.

CloudMade (2009). *Cloudmade Makes Maps Differently*. Available at: http://maps.cloudmade.com/ [Accessed June 7, 2009].

Cox, C.W. (1976). *Anchor Effects and the Estimation of Graduated Circles and Squares*. The American Cartographer, **3(1)**, 65-74.

Cuff, D.J. & Mattson, M.T. (1982). *Thematic Maps: Their Design and Production*, New York: Methuen, 169pp.

DeLucia, A. & Hiller, D. (1982). *Natural legend design for thematic maps*. The Cartographic Journal, **19(1)**, 46-52.

Dent, B.D. (1990). *Cartography: Thematic Map Design Second Edition*., Dubuque, IA: William C. Brown, 448pp.

Harrower, M. (2003). Tips for Designing Effective Animated Maps. *Cartographic Perspectives*, **44**, 63-65.

Jolivet, L. (2009). *Characterizing maps to improve on-demand cartography - the example of European topographic maps*. In Proc. GISRUK17. Durham, UK: University of Durham.

Jolivet, L. (2008). *On-demand map design based on user-oriented specifications*. In Proceedings AutoCarto 08. Shepherdstown, WV: CAGIS.

Robinson, A.H. et al. (1995). *Elements of Cartography Sixth Edition*., New York, NY: Wiley, 688pp.

Sieber, R., Schmid, C. & Wiesman, S. (2005). *Smart Legend - Smart Atlas*. In Proceedings XXII International Cartographic Conference 2005. A Coruna, Spain.

Slocum, T.A. et al. (2009). *Thematic Cartography and Geovisualization Third Edition*, Upper Saddle River, NJ: Prentice Hall, 576pp.

Sutton, E., Medyckyj-Scott, D. & Urwin, T. (2007). The EDINA Digimap Service 10 Years On. *The Cartographic Journal*, 44(3), 268-275.

Wood, J. & Dykes, J. (2008). Spatially Ordered Treemaps. *IEEE Transactions in Visualization & Computer Graphics*, **14(6)**, 1348-1355.


## Biographies

***Jackie Clark*** *is a Web and Graphic Designer at EDINA User Support with industrial experience working with Cadburys and at The Glasgow School of Art and an MA in Product Design from the Royal College of Art.*

***Dr. Jason Dykes*** *is a Senior Lecturer at the giCentre, City University London undertaking applied and theoretical research in, around and between information visualization, interactive analytical cartography and human-centred design.*

***Dr. Fiona Hemsley-Flint*** *is a GIS Technician at EDINA Geodata Services with interests in spatial databases, cartography, web mapping, and web map and data delivery.*

***Dr. David Medyckyj-Scott*** *was Manager of Research and Geodata Services at EDINA until December 2009. He now works for Landcare Research in New Zealand. His interests include geospatial metadata and portals, data sharing, web mapping, interoperability, Spatial Data Infrastructures and e-Research tools.*

***Lasma Sietinsone*** *is a GIS Technician at EDINA Geodata Services with interests in using open source technologies to provide web map and data delivery.*

***Dr. Aidan Slingsby*** *is a Willis Research Fellow at the giCentre at City University London with research interests in designing, implementing and using geovisualization techniques for assessing data quality and variability and for visual data analysis.*

***Tim Urwin*** *is Geospatial Data Manager at EDINA and responsible for the management of the Digimap Service. His interests are in spatial databases, cartography, web mapping and web map and data delivery.*

***Dr. Jo Wood*** *is a Reader in geographic information at the giCentre at City University London with research interests in geovisualization, terrain modelling and object oriented programming for spatial sciences.*

# Multi-Scale Visualisation of Inbound and Outbound Traffic Delays in London

Tao Cheng[1] Andy Emmonds[2] Garavig Tanaksaranond[1] Damilola Sonoiki[1]

[1]Department of Civil, Environmental and Geomatic Engineering,
University College London, Gower Street, WC1E 6BT London, UK
[2]Road Network Performance & Research,Transport for London,
197 Blackfriars Road, London SE1 8NJ, UK
{tao.cheng; g.tanaksaranond }@ucl.ac.uk; andy.emmonds@tfl.gov.uk; dami.sonoiki@googlemail.com

KEYWORDS: multi-scale; visualisation; traffic delay; inbound; outbound

## 1. Introduction

Large cities throughout the world are increasingly crowded as a result of increased population and mobility. Traffic congestion severely affects people's daily life and has enormous economic impact. The annual cost of congestion on London's main roads is estimated to be more than £3 billion. In order to maximise efficiency of people movement and to minimise delays and disruption in the transport system, we need to investigate and measure the performance of road networks and the impact of interventions on road capacity. In this way, traffic congestion can be effectively alleviated. To achieve this, it is very important to identify and understand the nature of the travel time delays experienced on the city's road network. It is believed that the increase in congestion reflects travellers' responses, both temporary and longer-term, to the scarcity of space of the road network. In order to gain in-depth understanding of the traffic patterns on a road network, there is a need to understand both the spatial and temporal characteristics of the road network.

Geovisualisation can not only help us to understand spatio-temporal variations in performance of the road network, but it will also help us to uncover hidden patterns of travel time relationships and delays occurring on the network. Attempts to generate travel time maps to analyse traffic related information date back to the 1960s. For examples, Bunge (1960) used irregular isochrones (lines of equal travel time) on a map to portray travel times from a given origin. Marchand (1973) used multidimensional scaling (MDS) to map travel times – specifically using average speeds achieved on different types of roads to estimate the travel times between cities in 1936, 1941, 1950 and 1961. Weir (1975) illustrated the use of maps in transportation planning by analyzing the changes in travel times arising from new road construction. Johnson (1981), Gatrell (1983) and Aitkin (1974), all used Q-Analysis to show relative change in travel times. These maps were however not easy to understand (Ahmed and Miller, 2007). Inoue et al. (2006) used travel time contour maps and time-space maps to view hourly change of the roadway level of service in Tokyo City using probe vehicle data. Wood et al. (2008) used treemaps to produce road maps of traffic volume and speed. Yudong et al. (2008) used self organizing maps (SOM) to analyse and visualise traffic flow time series in an urban traffic network. Google Maps (maps.google.com) have been updated to show historic and current traffic speed on the highways outside Central London as a bi-directional thematic map.

Despite the efforts that have been expended in representing travel time, there has to date been no serious attempt to represent traffic delays (excess travel time). Furthermore, previous research has also usually used a single visualisation technique and a single spatial or temporal scale. This paper explores the use of different geovisualisation techniques to produce travel time maps that show excess travel time delay in Central London. We generate travel time maps using two thematic views and two temporal intervals in order to show travel delays.

## 2. Data

The data, supplied by Transport for London (TfL), were captured as part of the current London Congestion Analysis Program (LCAP). Processed (i.e. post outlier removal) ANPR (Automatic Number Plate Reading) journey times are captured at five minute intervals for 539 core road links in Central London. The recorded journey time is the difference between the time recorded at the first node of a road link and the recorded time at the end node of the same road link. The data captured in April 2009 was supplied by TfL in 4 separate weekly data tables. Each table contains approximately 2.9 million rows of data. The data are bi-directional with inbound and outbound directions: this represents a difficult challenge in visual representation, in addition to the large volume of data.

The concept of excess travel time or travel time delay was adopted in order to adequately represent the true traffic conditions on each individual road. Baseline journey time is the average journey time obtained during free flow period and this was chosen to be between midnight and 6am. The average baseline journey time subtracted from the average journey time for each road link is the excess travel time for any particular road link. The calculated excess travel time for each road link has been aggregated 15 minute and hourly intervals.

## 3. Visualisation of travel time delays in Central London

We present traffic congestion patterns of excess travel times at two temporal scales and using two types of visualisation. Using a contour map, delay patterns are revealed at hourly intervals, while thematic maps are used to show the patterns at 15 minutes interval. The thematic map considers bi-directional movement in order to show both inbound and outbound directions (similar to the Google style).

The overall application is developed within the ESRI ArcGIS environment. The travel time delay contour maps provide a general overview of the overall delay conditions on all of the road networks. Excess travel time, as modelled in the underlying database, is selected by the user for a specific date, and hour of the day, and IDW is used to interpolate between nodes of the road links in order to generate contour maps that highlight areas likely to be experiencing congestion conditions. They show general traffic conditions at hourly intervals and can show inbound and outbound roads separately using the related directional data (see Figure 1).

The inbound contour map in the Figure 1 shows a larger area of traffic congestion than outbound (the inbound figure shows larger red and yellow area in the city center than the outbound figure). It also shows two highly congested zones in the morning peak: the red zone on the right in Barking and Dagenham and the red zone on the left near Kingston Bridge. The outbound contour map shows that there was a big red zone on the right (near centre) which is around the Southwark, Greenwich and Lewisham area. Another big red area in the right corner of the map was near Swanley.

The travel time thematic map provides greater details of the delays experienced on individual roads (Figure 2 and 3). The areas of congestion in the Figure 1 (red areas) can be clearly seen in Figure 2. For example, the large red area on the right hand side in the Figure 1 inbound is also shown clearly in Figure 2 that A13 was congested. Similar to Figure 1, Figure 2 also shows the difference between inbound and outbound roads. Figure 3 shows the delay at two different times of the day with bi-directions on one map. It is clear that in some links inbound and outbound direction are different. This figure also shows the delays in the city centre are changing in within day - the delays in the city center at the afternoon peak hour was severer than at the morning peak hour.

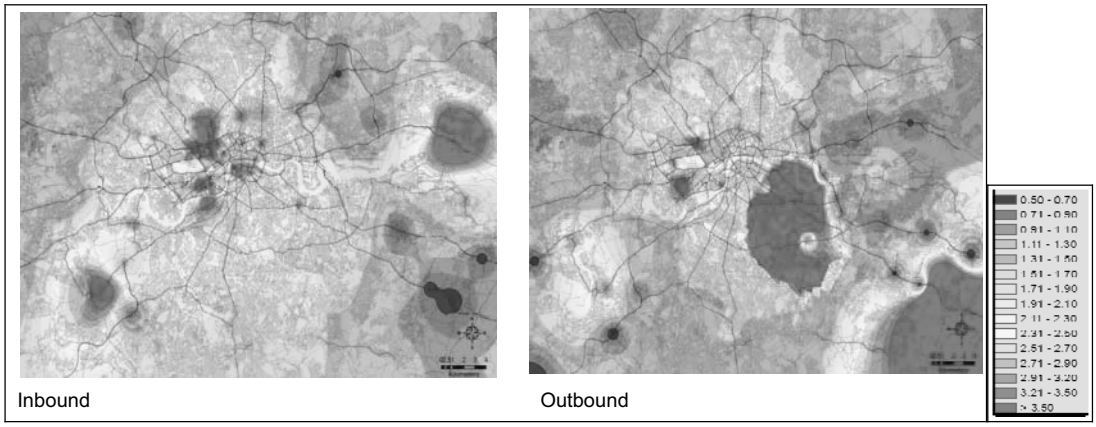



Figure 1. Overall travel delay (in minutes) at 9am on 12th April 2009

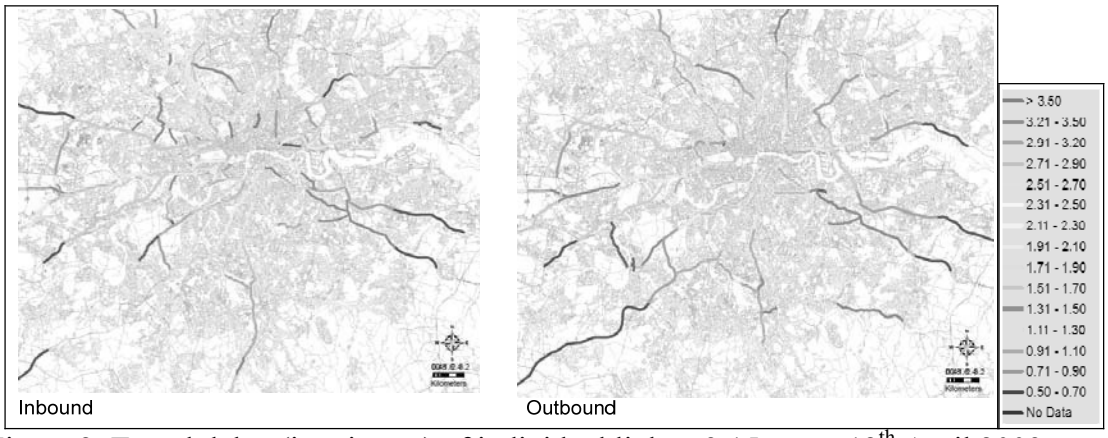



Figure 2. Travel delay (in minutes) of individual link at 9:15am on 12th April 2009

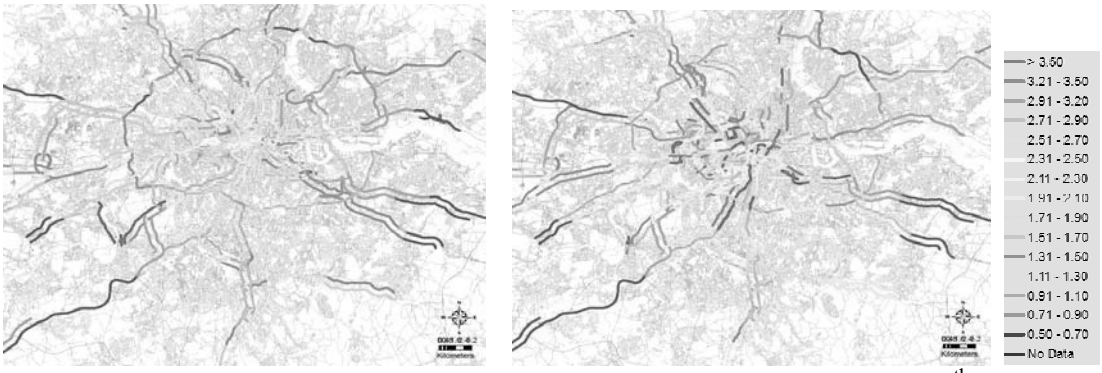



Figure 3. Bi-directional travel delay (in minutes) at 9:15am and 3:45pm on 12th April 2009

## 4. Conclusions and Discussion

Detecting patterns and understanding variations in travel times in a large dataset can be very complex. This study shows the travel time delays and the traffic conditions on the London road network at a number of spatial and temporal scales which complement each other. The contour maps show the general conditions and trip pattern distribution on the entire road network, while the thematic maps show the detailed pattern and delay condition of individual links. Visualising travel time delays at different scales can improve the user's cognitive understanding of prevailing traffic conditions and traffic commuting patterns on the road network. The exploration of travel time data using the proposed visualisation techniques allows the transport analyst to visually recognise and understand the geographical patterns of delay on different road networks, in order to facilitate further investigations. We are working

on 3D visualization of traffic delays in order to provide another perspective on traffic patterns in Central London. We believe the methodology developed here is applicable to other big cities around the world.

**Acknowledgements**

This research is support by UK EPSRC (EP/G023212/1) and Chinese NSF (40830530).

**References**

Aitkin, R.H., (1974). *Mathematical Structure in Human Affairs*. Heinemann, London.

Ahmed, N. and Miller, H. (2007). Time–space transformations of geographic space for exploring, analyzing and visualizing transportation systems. *Journal of Transport Studies* **5(1)**, pp 2-17.

Bunge, W. (1960). *Theoretical Geography*. Ph.D. dissertation, University of Washington.

De Oliveira, MCF and Levkowitz, H. (2003). From visual data exploration to visual data mining: a survey, *IEEE Transactions on Visualization and Computer Graphics* **9(3)**, pp 378–394.

Gatrell, A. (1983) Distance and space: a geographical perspective. Clarendon Press: Oxford.

Goldsberry, K. (2008) GeoVisualization of Automobile Congestion, *Proceedings of the AGILE 2008 Conference: GeoVisualization of Dynamics, Movement, and Change*. Girona, Spain, May 5, 2008.

Inoue R. (2006). Visualization of road travel time in Tokyo city center using probe vehicle data, http://planner.t.u-tokyo.ac.jp/member/inoue/pdf_files/UPIMap2006_probevehicle.pdf [access on December 7 2009].

Johnson, J.H. (1981). The Q-analysis of road traffic systems. *Environment and Planning B*, 8, 141–189.

Karlin, O. and McLaren, R. (2005), http://www.oskarlin.com/2005/11/29/time-travel/ [access on December 7 2009].

Keim, D. A. (2004). Visual data-mining techniques, in *Visualization Handbook*, C. Johnson and C. Hansen (eds), Elsevier Science Publishing, pp 813–825.

Lightfoot, C. and Kelly, F. (2009). Travel Time Maps. http://www.mysociety.org/projects/travelmaps/ [accessed on November 17 2009].

Marchand, B. (1973). Deformation of a transportation space. *Annals of the Association of American Geographers* **63 (4)**, pp 507–522.

Weir, S., (1975). *Getting Around Town: Modifications in a Local Travel Time Space Caused by Expressway Construction*. MS Thesis, Pennsylvania State University.

Rubin, J. (1994). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. New York: John Wiley & Sons.

Wood, J., Slingsby, A., Dykes, J. (2008). Using treemaps for variable selection in spatio-temporal visualization. *Proceedings of the AGILE 2008 Conference: GeoVisualization of Dynamics, Movement, and Change*. Girona, Spain, May 5, 2008.

Yudong C. (2008). *Multi-dimensional Traffic Flow Time Series Analysis with Self-Organizing Maps*. Tsinghua Science and Technology ISSN 1007-0214 16/20, 13, pp 220-228.

Tao Cheng is a senior lecturer in GeoInformatics in University College London. Her research interests span network complexity, integrated spatio-temporal data mining, spatial-temporal data modelling and visualisation, and uncertainty and quality of geographic information, with applications in coastal zone management, environmental monitoring, epidemics and transport studies.

Garavig Tanaksaranond is currently an MPhil/PhD student in Geomatics Engineering at University College London. She received a BEng in Survey Engineering and MSc in Statistical Information System from Chulalongkorn University in Thailand. Her research interests include geographical information science, visualization, and spatial databases.

Andy Emmonds is currently the Traffic Directorate and the Principal Transport Analyst of network performance in Transport for London (TfL).

Damilola Sonoiki holds an MSc degree in Geographic Information Science from University College London. After graduation, he kick started the Climate Change Adaptation mapping project at the London Borough of Camden. His research interest is currently looking into how artificial intelligence and natural language processing techniques can improve information sharing and interoperability.

# Preliminary Results of a Spatial Analysis of Dublin City's Bike Rental Scheme

Peter Mooney[1,2], Padraig Corcoran[1] and Adam C. Winstanley[1]

[1]Geotechnologies Research Group, Department of Computer Science, National University of Ireland Maynooth (NUIM), Maynooth, Co. Kildare. Ireland Tel. +353-1-708 3847
{peter.mooney, padraig.corcoran, adam.winstanley} @nuim.ie
[2]Environmental Research Center, Environmental Protection Agency, Richview, Clonskeagh, Dublin 14. Ireland p.mooney@epa.ie Tel: +353 (0)1 268 0181
KEYWORDS: Bike Rental, Spatial Analysis, Environmental Transport

## 1. Introduction

We present some initial observations on the usage and flow patterns of the DublinBikes (DB) bicycle rental scheme across Dublin city. In September 2009 Dublin City in conjunction with outdoor advertising company JC Decaux made 450 bicycles publicly available from 40 locations around the city in a scheme called DublinBikes (DB). Cycling, as a commuting mode forms an important part of the Irish Government's Transport policy for Ireland up to 2020 stating that "*a culture of cycling will be developed by 2020 to envisage around 160,000 people cycling for their daily commute, up from 35, 000 in 2006* " (DOT, 2009). We follow Froehlich et al (2008) who find usage patterns from these bike rental schemes can "*infer cultural and geographical aspects of the city and predict future bike station usage behaviour"* when combined with geographical information and local knowledge. Data captured on DB and presented in this paper covers the period of September 20th 2009 to February 15th 2010 inclusive.

### 1.1 Data Capture and Experimental Setup

There are 40 DB terminals or stations in Dublin with 450 DB bicycles available at full availability. Each DB terminal has between 15 and 25 spaces. Real time information (in XML format) is available from http://www.dublinbikes.ie on each DB terminal including information on the number of bikes available and the number of parking spaces currently available. We do not have access to the individual movements of DB bikes from checkout terminal to return terminal. The DB network is not fully self-supervising. In a self-supervising network some DB stations would suffer from unbalanced checkouts and returns of bikes. Some load balancing is performed by DB staff moving bikes between stations. An assumption in our work is that the forced load balancing only happens sporadically and consequently does not significantly bias our statistical results. The frequencies of bike checkouts and returns correspond to a terminal being characterised as "busy" or "not busy" (Froehlich et al, 2008). We use the OpenStreetMap database for Dublin to provide us with access to spatial data on locations of bus stops, train and metro stations, and other aspects of Dublin's transportation infrastructure.

## 2. Discussion of current results

The spatial layout of a city has an obvious influence on the movement patterns and social behaviours found within (Froehlich et al, 2008). Transportation systems providing good access to all transportation modes have a positive influence on movement patterns in a city (Brons et al, 2009). There are subtle differences between the patterns of bike checkouts and returns for all of the 40 stations. Currie (2009) outlines some reasons for these including social needs, population density, and public transportation service level.

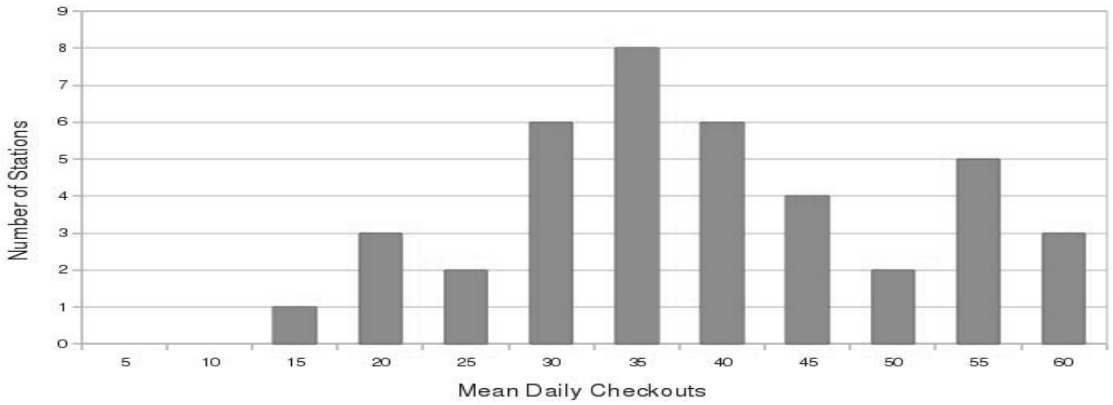



*Figure 1. Distribution of mean daily checkouts for all DB station terminals*

Figure 1 shows a frequency distribution of the mean number of checkouts per day for all DB terminals over the entire observation period. Three clusters of stations are immediately apparent – stations with a daily checkout mean of less than 25, stations with mean daily checkouts of between 25 and 45, and finally stations with mean daily checkouts of between 45 and 60 checkouts.

*Figure 2. Location of all 40 DB terminals. Larger circles indicate busier DB terminals. OpenStreetMap for Dublin is used as the base map*

Figure 2 shows the location of all 40 DB terminals. The largest circles indicate the location of the busiest terminals (combining checkouts and returns) while smallest circles indicate the location of the quietest terminals in terms of checkouts and returns of bikes. There is no obvious spatial clustering of the busiest terminals but attention is drawn to the three busy terminals at the southern portion of the map. These are located on a very busy orbital transport route in the city known as the "canal route". There is a higher concentration of terminals with low frequencies of checkouts and returns in the northern part of the city.

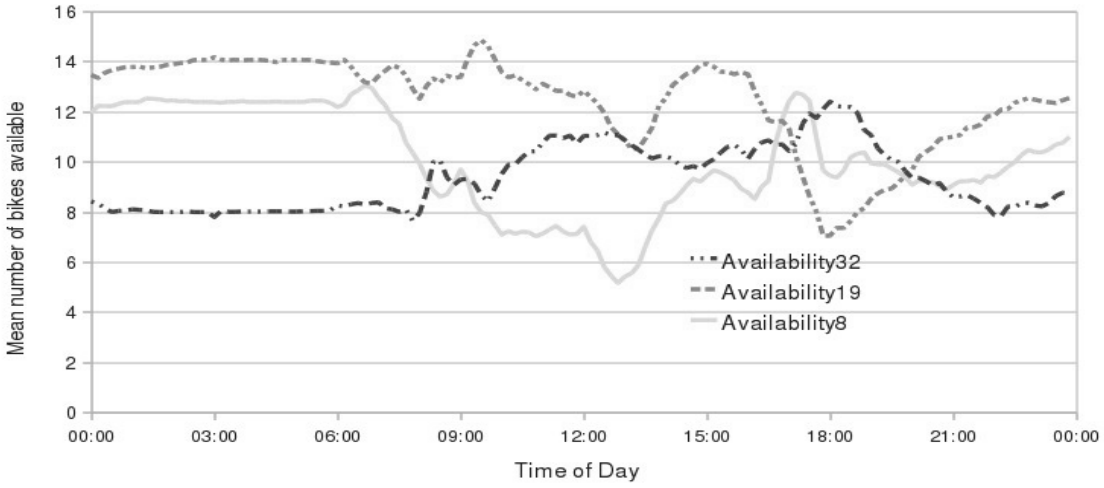



*Figure 3. Time-series plot of the mean number of bikes available at the three busiest stations in the DB network*

Figure 3 shows the mean number of bikes available at the three busiest stations in the DB network for all weekdays over the observation. Station 8 (denoted as Availability8) shows low availability during the working day. This could be linked to its location at a busy pedestrian bridge over the River Liffey beside the international financial center. Station 32 shows increasing availability of DB over the day. This DB terminal is beside Pearse Street mainline and commuter train station. This could indicate that people are using DB to move from other locations and park their bikes at Station 8 to possibly link to public train transport. Finally Station 19 is located in the suburbs. The dramatic decrease in availability at evening rush hour could indicate the movement of people away from this location as businesses close for the evening in this area.

Table 1 below shows statistics from a selected number of stations. These stations were chosen from the 40 stations for the purposes of illustrating the different characteristics for stations during the week and at weekends. For each station their mean daily checkouts from weekdays (CKWeek) and weekends (CKWend) are shown. The overall ranking of all stations based on CKWeek is shown in the column Pweek while the overall ranking of all stations based on CKWend is shown in column Pwend. The column Diff indicates if a particular station changes overall ranking from weekday to weekend. For example the first row (Pearse Station) is the busiest bike terminal (ranked 1$^{st}$) on weekdays with 49.67 checkouts per day but is only the 11$^{th}$ busiest overall at the weekend. The Diff column is -10 indicating a drop of 10 in overall ranking from weekday to weekend and consequently a drop in mean checkouts relative to all other stations. Some important observations can be made from the table above. The four busiest stations (32,8,19,5) during the week lose four or more rankings in overall mean checkouts at weekends. The most dramatic set of changes is for stations (20,25,22) dropping very significantly in overall ranking. When one looks at the locations of these stations they are within key office and business areas of Dublin city. It is no surprise to see the usage of DB terminals in these locations decrease at weekends when these areas become dramatically quieter. Stations (40,10,24) display the opposite effect with dramatic increases in overall ranking at weekends. These stations are located close to key shopping areas and leisure facilities in the city which naturally see a large increase in visitors at weekend. It is worth noting that station 15 is placed bottom of both the weekday and weekend ranking which is possibly due to its isolated location in the north inner city and close proximity to a high frequency bus corridor.

*Table 1. Comparison of mean checkouts for selected stations during weekdays and weekends*

| StationID | Pweek | CKWeek | Pwend | CKWend | Diff |
|---|---|---|---|---|---|
| **32 (Pearse)** | 1 | 49.67 | 11 | 9.35 | -10 |
| **8 (Custom Hse)** | 2 | 46.39 | 6 | 11.00 | -4 |
| **19 (Herbert)** | 3 | 43.59 | 7 | 10.21 | -4 |
| **5 (Charlemont)** | 4 | 43.21 | 12 | 9.30 | -8 |
| **11 (Earlsfort)** | 10 | 38.01 | 36 | 3.15 | -26 |
| **20 (James St. East)** | 11 | 38.00 | 37 | 2.98 | -26 |
| **25 (Merrion Sq East)** | 13 | 35.55 | 26 | 5.47 | -13 |
| **22 (Townsend Street)** | 18 | 29.72 | 30 | 4.63 | -12 |
| **40 (Jervis Street)** | 23 | 26.25 | 9 | 9.86 | +14 |
| **10 (Dame Street)** | 29 | 23.93 | 15 | 7.38 | +14 |
| **24 (Cathal Brugha)** | 33 | 21.67 | 22 | 5.67 | +11 |
| **15 (Hardwicke)** | 40 | 9.82 | 40 | 1.81 | 0 |

## 3. Discussion and Future Work

Initial analysis of the checkout and return statistics for each terminal appear to loosely support the findings of Martens (2004) who showed that in the Netherlands the closer bike parking stands and facilities were to bus/railway station entrances/exits the higher the use of bicycles as part of the access or exit trip to the station. During weekdays three of the busiest 5 stations (Four Courts, Pearse, Charlemont) are within 400 meters of either stations on the LUAS metro system or mainline/sub-urban train stations. However this changes at weekends where the 5 busiest stations are located with 400 meters of shopping centers and key shopping areas. Martens (2007) concludes that "bicycle usage in trips to and from public transportation and leisure facilities can be promoted simply by providing more sufficient and attractive bicycle parking facilities". At the time of writing 1,500,000 records are currently stored in the PostGIS database for the DB activity data. The analysis above has shown that there are patterns developing at a station level – the checkout and check-in of bikes and at the network level where stations close to major transportation locations are busier than those further away. Some initial patterns are developing where DB terminals are close to bus stops where the service level frequency of buses is high. The database of DB information is time-series data and we are currently investigating methods for similar time-series pattern retrieval. We have looked at usage patterns in DB by clustering DB terminals into geographically relevant groups – DB terminals: near train stations, within 300 meter walking distance of each other, at major street intersections. DB terminals without explicit geographical relationships such as proximity may exhibit similar time-series patterns. Quantifying how similar the time-series for non-geographically adjacent stations are may give us an insight into other aspects of the characteristics of the DB bike flows. Given the high dimensionality of the DB time series it may be necessary to reduce the dimensionality of the time-series for each DB terminal before attempting to perform a similarity search.

**References**

Benenson, Itzhak and Martens, Karel and Birfir, Slava. **PARKAGENT: An agent-based model of parking in the city**, Computers, Environment and Urban Systems, Vol 32(6), GeoComputation: Modeling with spatial agents, November 2008, Pages 431-439

Brons, Martijn and Givoni, Moshe and Rietveld, Piet. **Access to railway stations and its potential in increasing rail use**, Transportation Research Part A: Vol 43(2),2009, pp 136-149

Currie, Graham. **Quantifying spatial gaps in public transport supply based on social needs**, Journal of Transport Geography, In Press, Corrected Proof, Available online 2 May 2009

DOT, **SmarterTravel: Ireland's First National Cycle Policy Framework**. Department of Transport, Kildare Street, Dublin 2. Ireland. April 2009. Report available online at www.smartertravel.ie/

Froehlich, J., Neumann, J., and Oliver, N. (2009) **Sensing and Predicting the Pulse of the City through Shared Bicycling** Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), Pasadena, California, USA, July 11 - 17, 2009.

Martens, Karel. **The bicycle as a feedering mode: experiences from three European countries**, Transportation Research Part D: Transport and Environment, Vol 9(4), July 2004, pp 281-294

Martens, Karel. **Promoting bike-and-ride: The Dutch experience**, Transportation Research Part A: Policy and Practice, Volume 41, Issue 4, May 2007, Pages 326-338

**Biography**

*Peter Mooney is a postdoctoral researcher at the Dept of Computer Science and a Data Manager at the Environmental Protection Agency. Padraig Corcoran completed a B.Sc. in Computer Science and Software Engineering and a Ph.D. in Computer Science in 2004 and 2008 respectively. He is currently working as a lecturer and researcher in the department of Computer Science at the National University of Ireland Maynooth. Adam Winstanley gained MSc and PhD degrees in Computer Science in 1987 and 1991 respectively. Currently, he is Senior Lecturer and Head of Department of Computer Science and Senior Research Associate of the National Centre for Geocomputation (NCG) at NUI Maynooth. He is a Co-PI of the LBS strand of the STRAT-AG project at NCG.*

# An ontological modelling of communications for an intelligent transport environment

Seong Kyu Choi[1]

[1] Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E 6BT
Tel. (+44) 20 7679 2741 Fax (+44) 20 7380 0453
Email: s.choi@ucl.ac.uk

KEYWORDS: ontology, intelligent transport systems, spatiotemporal relations

## 1. Introduction

Geographic information describes features and phenomena on the Earth's surface by using their location in geographic space (x,y,z) and time (Goodchild et al, 1999). It is therefore not surprising that Geographic Information Systems (GIS) have been applied widely in the area of transport management to represent transport activities, as they lend themselves to such descriptions (Goodchild, 2000). Recently, with the advancement of Information and Communication Technologies (ICT), road transport has been evolving towards Intelligent Transport Systems (ITS). Through the use of communication and sensing, ITS will ensure the improvement of safety, mobility, and productivity of road transport services with real-time information (Maccubbin *et al.*, 2008; Ohmori *et al.*, 2000). Dedicated Short Range Communication (DSRC) was designed to support short-to-medium range (up to 1 km) ad hoc wireless communication technology for fast moving vehicles (up to 100 mph) and road infrastructure in the 5.9 GHz band (Fernandes and Nunes, 2007).

Although the communication component is of great importance, however, studies of suitable smantic contents (information messages) over these communication channels are still limited. This paper proposes a geo-ontology model for vehicle-to-vehicle communication and vehicle-to-infrastructure communication based on DSRC technologies. It explores how situation-specific spatiotemporal information and knowledge on roads can be shared and how vehicles and road facilities can interact with each other to resolve local situations in an ITS setting.

## 2. An ontological approach

In the area of information systems, an ontology can be used to describe shared information and knowledge formally and precisely with logical constants and a set of statements. An ontology has to be machine-interpretable for shared and consensual knowledge and formal representation (Studer *et al.*, 1998). There are some ontological approaches that have been suggested for transport systems. For example, Lorenz *et al.* (2005) proposed an Ontology of Transportation Networks (OTN) as an encoding of Graphic Data Format (GDF), which is a standard for storing and delivering geographical data of transport and traffic applications. Hornsby and King (2008) described four types of motion relations – isBehind, inFrontoOf, driveBeside, and passBy – for moving vehicles in a road network situation. These relations were used to describe how in each case of relations between two vehicles conceptually different semantics can exist.

However, to support DSRC communication among vehicles and road infrastructure, they have to be described in a single framework so that their spatiotemporal relations may assist their communications. This paper focuses on developing an ontological model for resolving specific situations by describing spatiotemporal relations among vehicles and infrastructure in an ITS setting. A scenario was developed to represent a situation for the vehicle-to-vehicle communication, and it

was described in an ontology model. For the ontology model, Web Ontology Language (OWL), SPARQL[1], and SPARQL Inferencing Notation (SPIN) were used as an ontology language, an ontology query language, and an inferencing rule, respectively.

## 3. An ambulance scenario

Enhanced communication among vehicles and infrastructure may not only provide information beyond drivers' visibility but also have a positive influence on whole transport system's efficiency and safety. Take the situation in Figure 1 as an example, in which an ambulance (A1) and another vehicle (V1) are following their own route and the routes meet at an intersection. If A1 is in a hurry to get to a hospital and there are cars travelling towards the same intersection as A1 and subsequently sharing the same road segments, A1 can request priority over the other vehicles. Their interactions and movements are described in Figure 2 with time stamps.

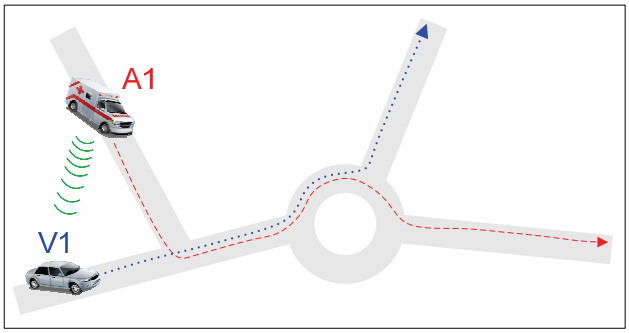
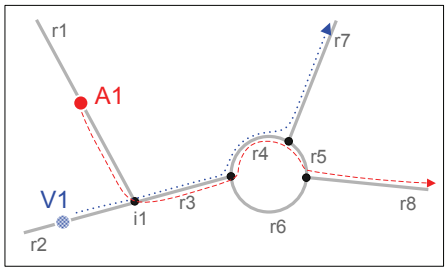
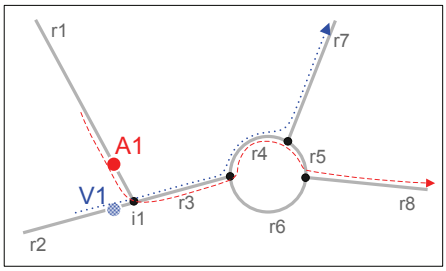


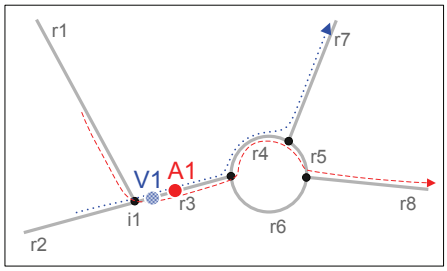

**Figure 1.** A scenario of an ambulance communicating with vehicles around it



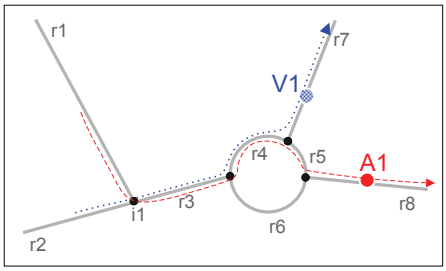

(a) A1 requests priority over cars towards r3.




(b) V1 slows down and stops at i1 to allow A1's priority.




(c) Both A1 and V1 are on r3. Now, A1 is in front of V1.




(d) Both A1 and V1 passed r3.

**Figure 2.** Snapshots of two vehicles' movements

[1] SPARQL is a recursive acronym that stands for SPARQL Protocol and RDF Query Language.

## 4. The VEIN ontology

In order to describe the above scenario in a single ontological framework, geographical objects such as road elements, intersections, roundabouts, and vehicles have to be classes of the ontology. The ontology may represent semantics in the scenario with classes, properties, axioms, instances, and the relationships (hierarchical relations and semantic relations) between them. Even though the above scenario describes vehicle-to-vehicle communication, vehicle-to-infrastructure communication is also important for DSRC implementation. Therefore, the VEIN ontology model was developed to conceptualize the two main categories (VEhicle and INfrastructure). As shown in Figure 3, upper classes for the ontology just started from GeoOWL. The geo:where property was used as an object property in which the domain is Feature class and the range is Geometry class (Lieberman *et al.*, 2007). The Feature class has three subclasses: VEhicle, INfrastructure, and Others. The Vehicle class indicates a mechanical road vehicle which has an engine and an On-Board Unit (OBU) for DSRC, while the Infrastructure class refers to road facilities which have a Road-Side Unit (RSU) for DSRC. VEhicle, INfrastructure, and Others may require more subclasses to be a domain ontology, and this is a work in progress.

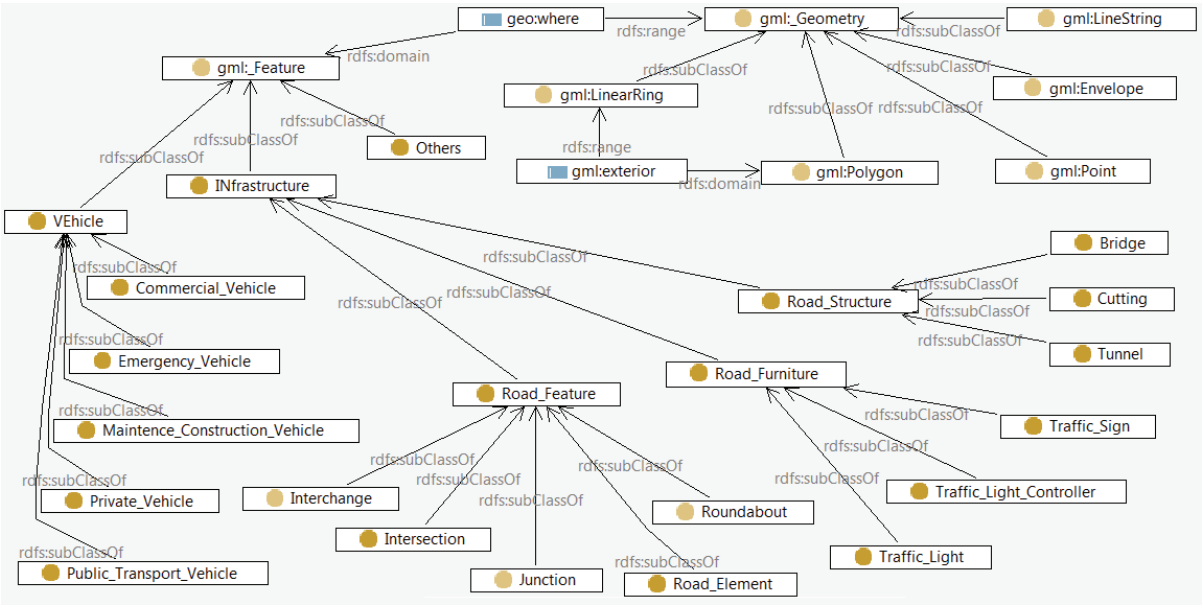



**Figure 3.** Classes of the VEIN ontology

**Table 1.** Object properties for spatiotemporal relations among vehicles and road elements

| Domain | Object property | Range |
|---|---|---|
| VEhicle | SpatialRelations:isLocatedBehind | VEhicle |
| VEhicle | SpatialRelations:isLocatedOn | Road_Element |
| Vehicle | Vein:nextRoadElementOfTheRoute | Road_Element |
| Road_Element | NetworkRelations:isConnectedTo | Road_Element |

The VEIN ontology was written in OWL, which was built on top of RDF/RDFS[2] in order to support

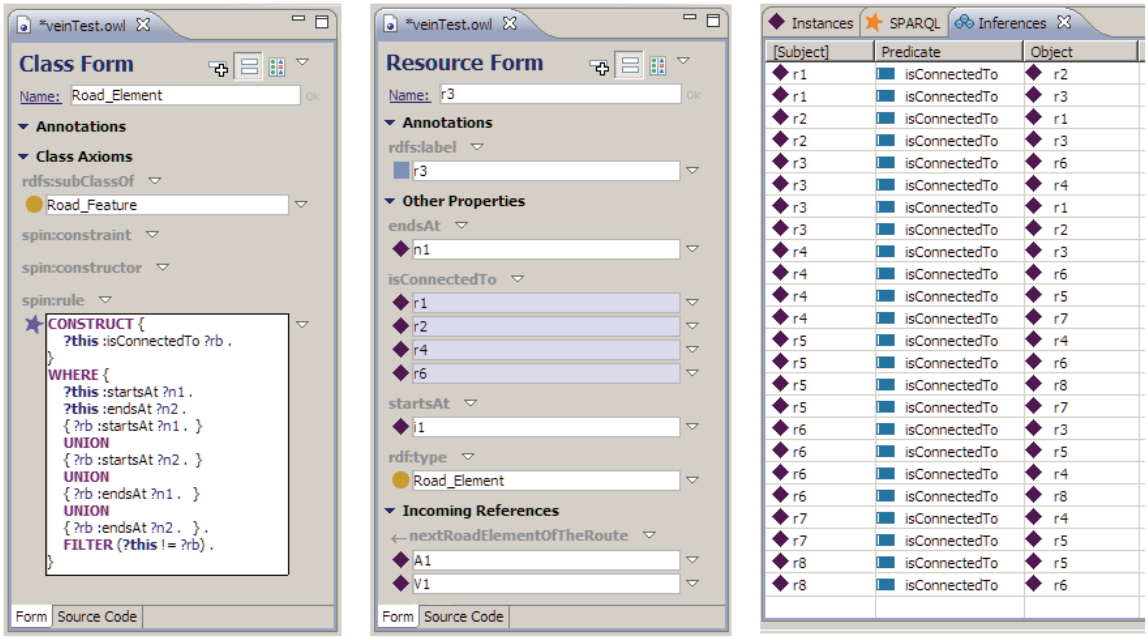
[2] Resource Description Framework (RDF) statements are triples (subject-predicate-object) to describe resources by Uniform Resource Identifiers (URIs), and a set of RDF triples forms a RDF graph. Additionally, RDF Schema (RDFS) semantically extends RDF and enables to specify classes of resources and their properties

extended vocabulary for cardinality constraints, richer property characteristics, etc. Some object properties in the VEIN ontology use those characteristics to describe spatiotemporal relations among vehicles and road elements. For example, in the Table 1, the object property SpatialRelations: isLocatedBehind is a transitive property and the inverse property of SpatialRelations: isLocatedInFrontOf. If there are some other geographical objects like traffic lights and traffic light controllers, more relations can be developed to describe the scene.

## 5. SPARQL to extract some information

Structured Query Language (SQL) is suitable for querying collections of tuples and is useful for extracting data from tabular and structured representations. In contrast, SPARQL is a query language for graph-based data so that it can traverse relationships easily and explicitly in RDF (Melton, 2006). While SPARQL has been designed and being used currently for querying RDF, it can be extended to OWL under the assumption that 'OWL 2 ontologies themselves are primarily exchanged as RDF documents' (OWL WG, 2009). Recently, SPARQL working group of W3C is trying to define the semantics of SPARQL queries for more detailed entailments including RDF Schema, OWL 2 Direct and RDF based semantics (Harris and Seaborne, 2010). In addition, SPARQL-based expressions can be used for inferencing and constraint checking in SPIN.

The property NetworkRelations:isConnectedTo in Table 1 is symmetric and express the connectivity between two road elements. Since every road element starts/ends at a junction, it can be constructed by SPIN. Figure 4 shows a SPIN rule in the Road_Element class to infer road elements' connectivity and its results. The SPIN rule is basically a SPARQL CONSTRUCT query, in which the system variable ?this, UNION, and FILTER were used for the current instance of the class, OR operation, and avoiding self-reference, respectively.




(a) An inference rule in the Road_Element class

(b) the (gray-coloured) values are generated automatically

(c) A set of new triples generated based on the rule

**Figure 4.** An inference rule and new triples for road elements' connectivity

In the above scenario, A1 requests priority over cars towards r3 (Figure 2a) and the set of vehicles in

(e.g. rdfs:Class, rdfs:Literal, rdfs:Datatype, rdfs:range, rdfs:domain, rdfs:subClassOf, rdfs:subPropertyOf, etc).

this situation that A1 has to communicate with can be divided into three parts (vehicles on r1, r2, and r3) depending on the connectivity of the road elements (Table 2). These queries can be executed if the location information of vehicles around A1 and their relations can be shared in the ontology. With these kinds of queries, A1 can minimize the number of vehicles it has to interact with at a time.

**Table 2.** SPARQL queries to support the scenario

| Purpose | SPARQL query | Result |
|---|---|---|
| To find vehicles in front of A1 on r1 | select ?car<br>where {<br>?ambulance rdfs:label "A1".<br>?ambulance SpatialRelations:isLocatedOn ?currentroad.<br>?car SpatialRelations:isLocatedOn ?currentroad.<br>?ambulance SpatialRelations:isLocatedBehind ?car.<br>} | none |
| To find vehicles on r2 (other roads connected to r3 except r1) | select ?car<br>where {<br>?ambulance rdfs:label "A1".<br>?ambulance SpatialRelations:isLocatedOn ?currentroad.<br>?ambulance :nextRoadElementOfTheRoute ?nextroad.<br>?currentroad NetworkRelations:isConnectedTo ?connectedroad.<br>?nextroad NetworkRelations:isConnectedTo ?connectedroad.<br>?car SpatialRelations:isLocatedOn ?connectedroad.<br>} | V1 |
| To find vehicles on r3 | select ?car<br>where {<br>?ambulance rdfs:label "A1".<br>?ambulance :nextRoadElementOfTheRoute ?nextroad.<br>?car SpatialRelations:isLocatedOn ?nextroad.<br>} | none |

## 6. Conclusion and future work

The proposed VEIN ontology allows for the semantic description of transport-related geographical objects and their spatiotemporal relations in specific situations. SPARQL queries were also built in order to assist the communications among vehicles and to minimize communication targets based on the query result extracted from the ontology. The VEIN ontology may support describing the ITS domain itself, and the implementation of the communications among vehicles and infrastructure.

With the ontology, vehicles and infrastructure can infer their context and semantics based on their spatiotemporal properties so that they can significantly improve the effectiveness of transport systems. Future work on this research will include the development of various scenarios to evaluate and redesign the VEIN ontology. In addition, there is a need to develop the mecahnism to share and update the ontology among fast moving vehicles. Also, an agent-based simulation will be implemented to evaluate the ontology, and to find the proper coverage and update interval of ontology queries in a dynamic vehicular environment.

## References

Fernandes, P., Nunes, U. (2007) VEHICLE COMMUNICATIONS: A SHORT SURVEY, IADIS

Telecommunications, Networks and Systems 2007 conference, Lisbon, Portugal.

Goodchild, M. F. (2000) GIS and Transportation: Status and Challenges, GeoInformatica, vol. 4, no. 2, pp. 127-139.

Goodchild, M.F., Egenhofer, M.J., Kemp, K.K., Mark, D.M. and Sheppard, E. (1999) Introduction to the Varenius project. International Journal of Geographical Information Science, 13, pp. 31–745.

Harris, S., Seaborne, A. (2010) SPARQL Query Language 1.1, W3C Working Draft 26 January 2010, http://www.w3.org/TR/2010/WD-sparql11-query-20100126/

Hornsby, K. S., King, K. (2008) Modeling Motion Relations for Moving Objects on Road Networks, GeoInformatica, Volume 12 (Number 4), pp. 477-495, Springer Netherlands.

Lieberman, J., Singh, R., Goad, C. (2007) Geospatial Ontologies, W3C Incubator Group Report 23 October 2007, http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/

Lorenz, B., Ohlbach, H. J., Yang, L. (2005) Ontology of Transportation Networks, REWERSE Deliverable A1-D4, http://idefix.pms.ifi.lmu.de:8080/rewerse/index.html#REWERSE-DEL-2005-A1-D4

Maccubbin, R. P., Staples, B. L., Kabir, F., Lowrance, C. F., Mercer, M. R., Philips, B. H., Gordon, S. R. (2008) ITS Benefits, Costs, Deployment, and Lessons Learned: 2008 Update, U.S. Department of Transport, Research and Innovative Technology Administration

Melton, J. (2006) SQL, Xquery, and SPARQL - What's Wrong With This Picture? XTech 2006: “Building Web 2.0”, Amsterdam, The Netherlands, 16-19 May 2006

Ohmori, S., Yamao, Y., Nakajima, N. (2000) The future generations of mobile communications based on broadband access technologies, IEEE Communications Magazine, Volume: 38, Issue: 12, pp: 134- 142

OWL Working Group (2009) OWL 2 Web Ontology Language, Document Overview, W3C Recommendation 27 October 2009, http://www.w3.org/TR/2009/REC-owl2-overview-20091027/

Studer , R., Benjamins, V. R., Fensel, D. (1998) Knowledge engineering: principles and methods. IEEE Transactions on data and Knowledge Engineering 25 (1-2):161-197

## Biography

*Seong K. Choi is a PhD student in the department of Civil, Environmental and Geomatic Engineering at UCL. He holds a BSc and a MSc in Geographic information Science and his research focuses on the interaction modelling and simulation based on spatiotemporal relations among geoobjects in ubiquitous computing environments.*

# GPS Data Collection Setting For Pedestrian Activity Modelling

Adel Bolbol[1], Tao Cheng[2]

[1&2]University College London UCL, WC1E 6BT
Tel. (+44)2076792781
a.bolbol@ucl.ac.uk, tao.cheng@cege.ucl.ac.uk

KEYWORDS: Epoch Rate, GPS, Activity Modelling, Spatiotemporal Behaviour, Transport Mode

## 1. Introduction

Modelling human's spatiotemporal behaviour has been an area of great research interest in the recent years. More recently a few research groups attempted to learn and deduce people's travel behaviour from sensor positional data (e.g. GPS, RFID…etc). This was done by deducing the mode of transport, route taken, start and end locations of trips and the purpose of the trip (Liao et al., 2005; Stopher, 2008). These studies usually use either online or offline sensors such as a GPS to collect positional data, where GPS measurements are usually made at a given frequency (e.g. every 30 seconds) or "epoch rate".

Although it might seem obvious that the more data the better, however, too much data will cause problems like increasing the processing time and the impracticability of battery requirements nowadays. Furthermore, the epoch rate chosen has a significant impact on the GPS data which results in errors in travel behaviour analysis such as "Cold Starts" (i.e. the device doesn't begin recording at the exact start/end location of a trip) or "Miss-Map Matching" (i.e. problems of lateral movement of position from the GPS trace) (Stopher, 2008).

There are attempts in finding the most suitable epoch rate for monitoring animals (Perotto-Baldivieso, et al. 2008), however, not much research give detailed analysis of the best GPS epoch rate setting for pedestrian activity modelling. This study attempts to find the best epoch rate setting on an offline GPS device in order to deduce the route and locations of the start and end of the trips. In broader terms, this study aims at providing the best configuration of a GPS device in order to have more meaningful data that arguably would lead to a better representation of human travel behaviours.

## 2. Data

The device used was a GPS device developed by u-Blox. This device has a positional accuracy of ±4.3 meters (u-Blox, 2009); however, the accuracy is significantly affected by the conventional GPS sources of errors in urban areas.

### 2.1. Data Specification

The Dataset selected was for around a 7-8 minute walking journey, which could be assumed to be the least time a pedestrian would walk on typical short journeys (e.g. from car-work, home-car, home-bus stop…etc) (*figure 1*).

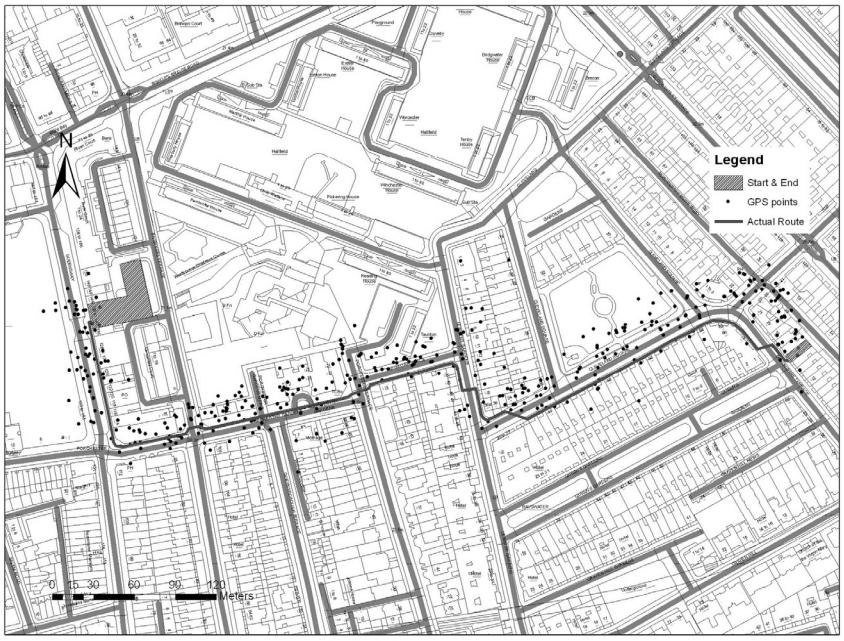



**Figure 1.** 1S Dataset with Actual Route

Data was collected every 1 second for the test track within London, starting at a source and ending at a final location. The data was then downloaded, processed and exported into ".csv" for further analysis.

The data was thinned in excel to the following epoch rates: 1, 10, 20, 30, 60, 120 and 300 seconds. The dataset was thinned 11 times in order to study the different probabilities from having 11 different datasets from every epoch rate group. From *figure 2*, it could be noted that at different thinning levels some epoch rates could lose some data due to lack of fixes at particular times.

| Dataset 1 - Thin 1 | | | | | | | | | Dataset 1 - Thin 2 | | | | | | | | | Dataset 1 - Thin 3 | | | | | | | | | Dataset 1 - Thin 4 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 20 | 30 | 60 | 120 | 300 | T | T+1 | 1 | 10 | 20 | 30 | 60 | 120 | 300 | T+2 | 1 | 10 | 20 | 30 | 60 | 120 | 300 | T+3 | 1 | 10 | 20 | 30 | 60 | 120 | 300 |
| 19.13 | | | | | | | 55 | 56 | 19.13 | | | | | | | 57 | 19.13 | | | | | | | 58 | 19.13 | | | | | | |
| 17.34 | | | | | | | 56 | 57 | 17.34 | | | | | | | 58 | 17.34 | | | | | | | 59 | 17.34 | | | | | | |
| 17.15 | | | | | | | 57 | 58 | 17.15 | | | | | | | 59 | 17.15 | | | | | | | 60 | 17.15 | 17.15 | 17.15 | 17.15 | 17.15 | | |
| 6.88 | | | | | | | 58 | 59 | 6.88 | | | | | | | 60 | 6.88 | 6.88 | 6.88 | 6.88 | 6.88 | | | 61 | 6.88 | | | | | | |
| 18.95 | | | | | | | 59 | 60 | 18.95 | 18.95 | 18.95 | 18.95 | 18.95 | | | 61 | 18.95 | | | | | | | 62 | 18.95 | | | | | | |
| 28.85 | 28.85 | 28.85 | 28.85 | 28.85 | | | 60 | 61 | 28.85 | | | | | | | 62 | 28.85 | | | | | | | 63 | 28.85 | | | | | | |
| 11.95 | | | | | | | 61 | 62 | 11.95 | | | | | | | 63 | 11.95 | | | | | | | 64 | 11.95 | | | | | | |
| 11.78 | | | | | | | 62 | 63 | 11.78 | | | | | | | 64 | 11.78 | | | | | | | 65 | 11.78 | | | | | | |
| 13.15 | | | | | | | 63 | 64 | 13.15 | | | | | | | 65 | 13.15 | | | | | | | 66 | 13.15 | | | | | | |
| 5.31 | | | | | | | 64 | 65 | 5.31 | | | | | | | 66 | 5.31 | | | | | | | 67 | 5.31 | | | | | | |
| 12.33 | | | | | | | 65 | 66 | 12.33 | | | | | | | 67 | 12.33 | | | | | | | 68 | 12.33 | | | | | | |
| 24.69 | | | | | | | 66 | 67 | 24.69 | | | | | | | 68 | 24.69 | | | | | | | 69 | 24.69 | | | | | | |
| 29.00 | | | | | | | 67 | 68 | 29.00 | | | | | | | 69 | 29.00 | | | | | | | 70 | 29.00 | 29.00 | | | | | |
| 21.31 | | | | | | | 68 | 69 | 21.31 | | | | | | | 70 | 21.31 | 21.31 | | | | | | 71 | 21.31 | | | | | | |
| 32.38 | 32.38 | | | | | | 70 | 71 | 32.38 | | | | | | | 72 | 32.38 | | | | | | | 73 | 32.38 | | | | | | |
| 23.13 | | | | | | | 71 | 72 | 23.13 | | | | | | | 73 | 23.13 | | | | | | | 74 | 23.13 | | | | | | |
| 22.54 | | | | | | | 73 | 74 | 22.54 | | | | | | | 75 | 22.54 | | | | | | | 76 | 22.54 | | | | | | |
| 6.75 | | | | | | | 81 | 82 | 6.75 | | | | | | | 83 | 6.75 | | | | | | | 84 | 6.75 | | | | | | |
| 11.43 | | | | | | | 83 | 84 | 11.43 | | | | | | | 85 | 11.43 | | | | | | | 86 | 11.43 | | | | | | |
| 11.84 | | | | | | | 85 | 86 | 11.84 | | | | | | | 87 | 11.84 | | | | | | | 88 | 11.84 | | | | | | |
| 24.93 | | | | | | | 87 | 88 | 24.93 | | | | | | | 89 | 24.93 | | | | | | | 90 | 24.93 | 24.93 | | 24.93 | | | |
| 0.42 | | | | | | | 88 | 89 | 0.42 | | | | | | | 90 | 0.42 | 0.42 | | 0.42 | | | | 91 | 0.42 | | | | | | |
| 3.49 | | | | | | | 89 | 90 | 3.49 | 3.49 | | 3.49 | | | | 91 | 3.49 | | | | | | | 92 | 3.49 | | | | | | |
| 10.41 | 10.41 | | 10.41 | | | | 90 | 91 | 10.41 | | | | | | | 92 | 10.41 | | | | | | | 93 | 10.41 | | | | | | |

**Figure 2.** Missing Data due to lack of Fixes in a Dataset

## 3. Method

### 3.1. Context

Considering the context of detecting the route and mode of transport, the most significant factors affecting such a probabilistic model would be the following:

- Route taken (Road detection)
- Distance travelled
- Trajectory speed
- Identifying Trip Start & End

Therefore, the following factors were studied in order to compare different epoch rates with respect to results from the previous list.

- Positional Errors
- Route Length Errors
- Average Speed Errors
- Distances from last points to Trips Starts & Ends

### 3.2. Calculations

*Figure 3* shows how the factors in the previous list (in yellow) are integrated with the conventional human behaviour model used in order to detect the transport mode and trip purpose. It illustrates how raw GPS data is snapped to the actual road links and then route length, speed and positional errors are calculated.

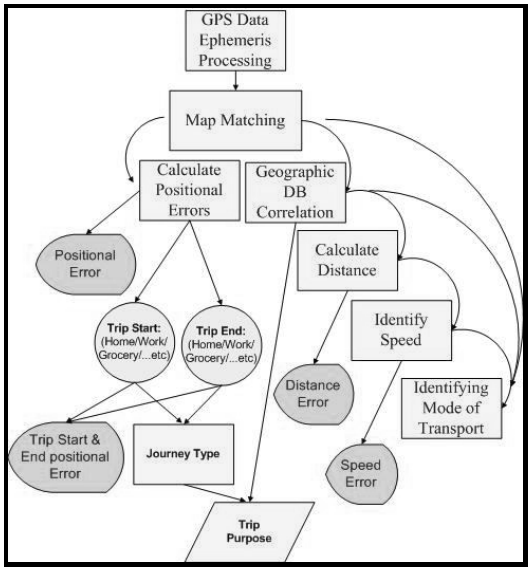



**Figure 3.** Study Investigations in Human Behaviour Modelling Context

The following is a description of how each of these processes is carried out:

- **Map Matching**

Map matching is first applied to all the thinned datasets. It is a method of snapping the GPS points to the correct road link. In a vehicle trajectory context a topological map matching algorithm would be a good choice since it is superior compared with other existing algorithms (Quddus, 2006). However, in a pedestrian-based trajectory, a better map matching technique would be a simpler algorithm using a point-to-curve approach together with a heading and speed taken into consideration.

- **Calculating Positional Error**

Horizontal positional error was calculated for each point from the actual route taken on the road network for each dataset.

- **Calculating Route Length Errors**

Euclidean distances were calculated between successive pairs of points for all points in each data set, along with the speed at each segment. The total length of the track is later calculated from each dataset and compared with the actual length of the track.

- **Calculating Average Speed Errors**

By the knowledge of the distances and time interval between each two records, the speed is determined for each segment, and the average speed along the whole route is calculated along with the standard deviation (δ).

- **Distances from last points to Trips Starts & Ends:**

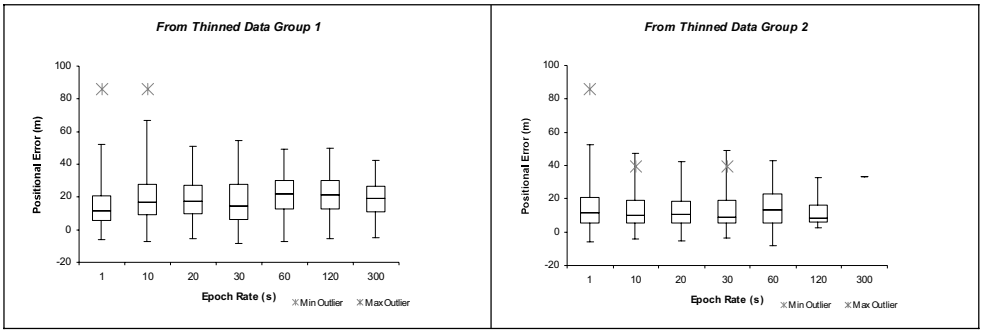
The aim of this test is to identify the start/end destination from where the GPS tracks start/end. This would depend on the last points in the dataset. Given that there are 11 datasets thinned from one, the last point will always change except for the 1s dataset. The distance between the last point from every dataset and the end destination is measured.


**Figure 4.** Actual Route with Data from Different Datasets near the End Destination

## 4. Analysis

### 4.1. Positional Error

Only 55 records out of 470 did not get a fix, and 2 out of 416 points had accuracy worse than 60m. The following box-plots illustrate the different positional errors of each of the 11 thinned datasets. The averages of all datasets tend to be close in value. There aren't many points in the 300s datasets and sometimes none; therefore, they prove to be not sufficient for data collection in that context. The 1s dataset on the other hand has more variances and instances which might badly affect the map matching process.



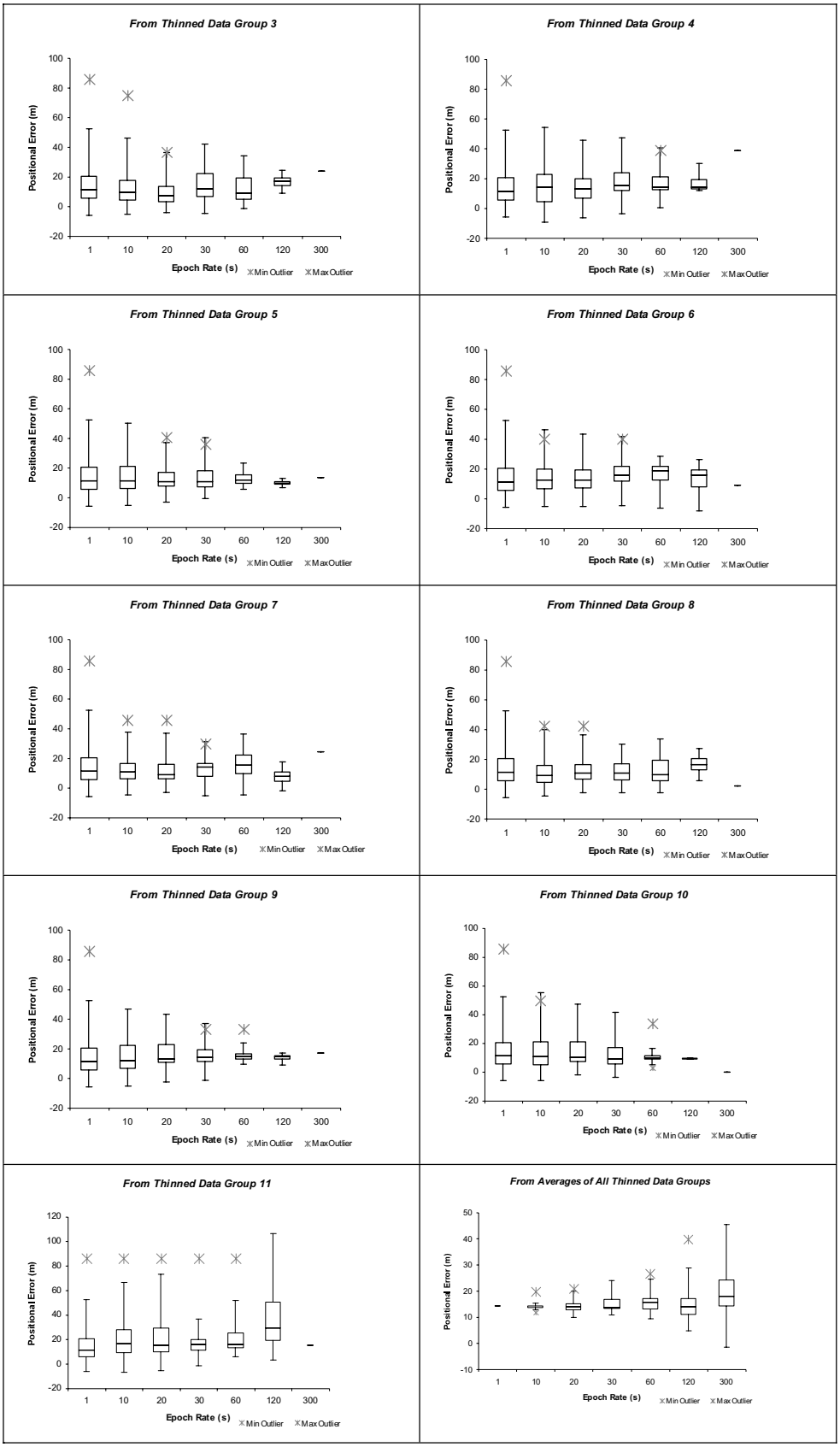



**Figure 5.** Box-plots Illustrating Positional Error in each of the 11 Thinned Datasets

### 4.2. Route Length Errors

Results for the length and speed are shown in table 1. As illustrated from *table 1* and *figure 6*; the route's lengths computed from 20s data prove to be the closest to the actual route length. Following that, are the 10s and 30s data, however, there is a higher certainty with the 30s due to closeness of its results. 1s results have been discarded from *figure 6* due to its high error value, for better representation purposes.

**Table 1.** Route Length & Average Speed from the 11 Datasets & Actual Route

| 1 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Length | | | | | | 4143.03 | | | | | |
| Speed | | | | | | 9.33 | | | | | |
| St Dev | | | | | | | | | | | |
| 10 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 952.52 | 766.07 | 834.73 | 829.03 | 739.29 | 690.01 | 659.70 | 723.44 | 793.99 | 879.54 | 952.52 |
| Speed | 2.16 | 1.72 | 1.93 | 1.78 | 1.49 | 1.50 | 1.50 | 1.61 | 1.90 | 1.98 | 2.16 |
| St dev | 2.25 | 1.78 | 2.02 | 1.77 | 1.47 | 1.58 | 1.80 | 1.66 | 1.86 | 1.69 | 2.25 |
| 20 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 643.32 | 680.95 | 595.59 | 746.85 | 585.13 | 615.05 | 619.02 | 611.57 | 679.19 | 622.64 | 773.65 |
| Speed | 1.42 | 1.59 | 1.37 | 1.70 | 1.28 | 1.44 | 1.41 | 1.47 | 1.59 | 1.38 | 1.63 |
| St dev | 0.93 | 1.05 | 0.81 | 1.21 | 0.92 | 1.02 | 0.88 | 0.88 | 0.74 | 0.92 | 1.58 |
| 30 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 600.31 | 580.52 | 585.92 | 580.40 | 555.01 | 580.08 | 576.59 | 587.96 | 538.81 | 586.80 | 754.10 |
| Speed | 1.35 | 1.43 | 1.42 | 1.37 | 1.34 | 1.40 | 1.36 | 1.40 | 1.32 | 1.33 | 1.72 |
| St dev | 0.70 | 0.57 | 0.48 | 0.61 | 0.55 | 0.49 | 0.52 | 0.56 | 0.78 | 0.56 | 1.04 |
| 60 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 543.57 | 466.81 | 484.19 | 484.14 | 495.79 | 493.86 | 500.40 | 492.45 | 480.82 | 493.65 | 683.02 |
| Speed | 1.29 | 1.33 | 1.37 | 1.34 | 1.38 | 1.37 | 1.37 | 1.37 | 1.34 | 1.33 | 1.63 |
| St dev | 0.38 | 0.36 | 0.33 | 0.34 | 0.31 | 0.25 | 0.27 | 0.23 | 0.26 | 0.15 | 0.63 |
| 120 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 422.33 | 315.72 | 178.35 | 315.72 | 315.72 | 313.69 | 180.55 | 178.35 | 339.63 | 180.55 | 490.98 |
| Speed | 1.17 | 1.32 | 1.49 | 1.32 | 1.32 | 1.31 | 1.50 | 1.49 | 1.42 | 1.50 | 1.36 |
| St dev | 0.29 | 0.17 | 0.00 | 0.17 | 0.17 | 0.20 | 0.00 | 0.00 | 0.07 | 0.00 | 0.28 |
| 300 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 298.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Speed | 0.9941 | | | | | | | | | | |
| St dev | | | | | | | | | | | |
| Actual | Actual Route Data | | | | | | | | | | |
| Length | 667.00 | | | | | | | | | | |
| Speed | 1.20 | | | | | | | | | | |

*Length (m), Speed (m/s)*

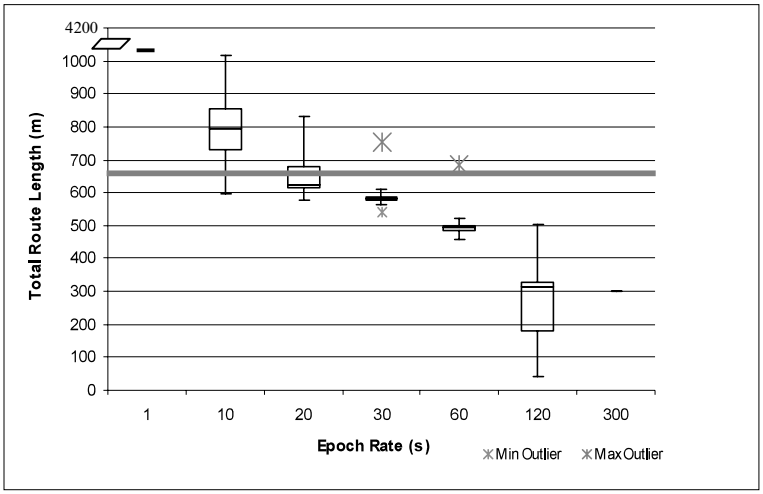



**Figure 6.** Total Route Length Computed from Datasets Compared to Actual Length

### 4.3. Average Speed Errors

Average speed results in *figure 7* demonstrate that 30s and 60s data give the best results and have a high level of certainty, followed closely by the 120s and 20s. 1s results were also discarded due to their high error value, while 300s has only one result because only one 30s thinned-dataset had 2 points.

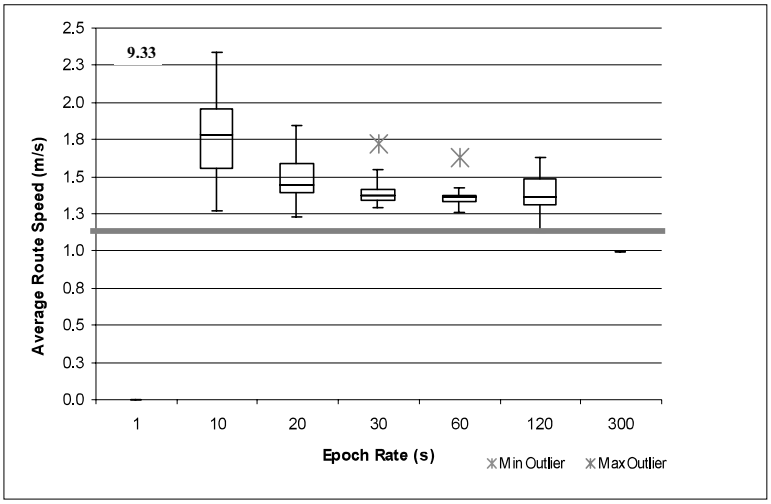



**Figure 7.** Average Speed Computed from Different Datasets Compared to Actual Average Speed

### 4.4. Distances from last points to Trips Starts & Ends

Since the 1s data is only 1 Dataset, the Start-End Trip-distance results are the same for all datasets, because only one point is chosen in all cases of data thinning. However, in any probabilistic model aiming at identifying the start/end location of a trip, the excessive data obtained from 1s epoch rate make it more likely to get better results but with high uncertainty. Still however, results from 10s and 20s data from *figure 8* and table 2 are good enough to find the start/end location, supported by the fact that GPS accuracy within an urban area is around 15-20m (Sec.5.1), and adding to that the length it takes a 10s or 20s-set GPS to get a fix is around 12m or 24m respectively (given speed=1.2m/s), would result in a distance of maximum 44m under the worst case scenario with 20s.

**Table 2.** Start-End Trip Distance Results

| | *1s* | *10s* | *20s* | *30s* | *60s* | *120s* | *300s* |
|---|---|---|---|---|---|---|---|
| Dataset1 | 1.6 | 1.6 | 1.6 | 1.6 | 45.3 | 116.8 | 210.0 |
| Dataset2 | 1.6 | 1.6 | 1.6 | 1.6 | 45.3 | 120.7 | 210.0 |
| Dataset3 | 1.6 | 8.0 | 1.6 | 8.0 | 45.3 | 120.7 | 210.0 |
| Dataset4 | 1.6 | 0.1 | 1.6 | 1.6 | 45.3 | 120.7 | 210.0 |
| Dataset5 | 1.6 | 8.0 | 8.0 | 1.6 | 45.3 | 120.7 | 198.3 |
| Dataset6 | 1.6 | 1.6 | 0.1 | 1.6 | 45.3 | 120.7 | 198.3 |
| Dataset7 | 1.6 | 1.6 | 1.6 | 8.0 | 45.3 | 120.7 | 187.3 |
| Dataset8 | 1.6 | 1.6 | 1.6 | 8.0 | 66.4 | 120.7 | 203.6 |
| Dataset9 | 1.6 | 1.6 | 1.6 | 66.4 | 66.4 | 120.7 | 196.3 |
| Dataset10 | 1.6 | 1.6 | 0.1 | 8.0 | 66.4 | 120.7 | N/A |
| Dataset11 | 1.6 | 1.6 | 1.6 | 1.6 | 1.6 | 1.6 | 193.6 |

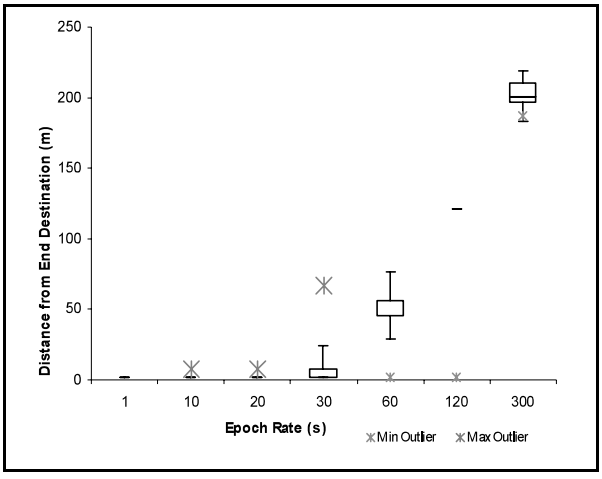



**Figure 8.** Distances from Set of Last Points in Each Dataset to the End Destination

## 5. Conclusions and Further Work

The study has shown that GPS data could have a mean accuracy of 15-20m within an urban area. In long epoch rate sessions, some bad quality data might influence the whole dataset to have a bad overall positional quality. However, a 1s dataset might too lead to bad map matching due to excessive scattered data.

The study has also shown that 20s data prove to be superior for route length calculation, while that 120s data gives bad accuracy in that context. It also shows that 30s and 60s data give the best accuracy for calculating the pedestrian speed. And for determining the start/end of a trip, the 10s and 20s data have shown the best results, bearing in mind that 1s data will also give a good accuracy but with high uncertainty.

And as a whole, the study concludes that the most appropriate epoch rate for route and start/end detection of pedestrians is probably either 20s or 30s sessions according to this study's datasets. This could arguably increase the accuracy of travel mode and route detection of GPS users. Also, more research will be done to conduct a similar study for other transport modes.s

## 6. Acknowledgements

We would like to express our gratitude to u-Blox & EPSRC, the sponsoring bodies for this research, as well as UCL for providing the facilities and research resources, and Dr. Chris Marshall & Andy Yule for their invaluable ideas and support. Also a special thanks to Ramathan Ali.

## References

Liao L, Fox D and Kautz H (2005). Location-based Activity Recognition using Relational Markov Networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp 773–778

Perotto-Baldivieso HL, Cooper SM, Figueroa-Pagan M, and Romo J (2008). Too Much Data? Optimizing GPS Collar Data Collection Schedules [poster]. *The 2008 Joint Meeting of the Society for Range Management and the America Forage and Grassland Council.* Louisville, KY: 30 January, 8:00 am-5:00 pm.

Quddus MA (2006). *High integrity map-matching algorithms for advanced transport telematics applications*, PhD Thesis. Centre for Transport Studies, Imperial College London, UK.

Stopher PR (2008) Collecting and Processing Data from Mobile Technologies, Resource Paper, *8th International Conference on Transport Survey Methods*, Annecy, France, May, CDROM

u-Blox (2009) YUMA: Software and Service for Capture & Process. Available at: *http://www.u-Blox.com/en/gps-solutions/yuma.html* [Accessed 04 October 2009]

**Biography**

*Adel Bolbol is a PhD student of GeoInformatics in the Civil, Environmental& Geomatics Engineering department, University College London. Recently, he studied at City University, London, pursuing a MSc in GIS in the Informatics department. He also has over 7 years of industrial expertise in Geomatics, GIS, HCI, Project Management, Quality & Environmental Management Systems. Adel's research interests and expertise include GIS, HCI, GeoVisualization, LBS and GPS.*

*Tao Cheng is a senior lecturer in GeoInformatics in the University College London. She has studied and lectured in China, the Netherlands, Hong Kong, France and the UK. Tao Cheng's research spans spatial-temporal modelling, analysis and visualisation, uncertainty and quality of geographic information, and spatio-temporal data mining and knowledge discovery. She has over 90 publications and she received The U. V. Helava Award for the Best Paper in the ISPRS Journal of Photogrammetry and Remote Sensing in the year 2000.*

# Revealing the Fuzzy Geography of an Urban Locality

Richard Flemmings

Department of Geography, Environment and Development Studies, Birkbeck, University of London, Malet Street, London WC1E 7HX
Tel. +44(0)207 631 6473 | Email: richflemmings@gmail.com

KEYWORDS: Geo-tagging, fuzzy, vernacular, kernel density estimation

## 1. Introduction

Urban localities are a hidden geography. They have no officially defined boundary. Boundaries of urban localities are subjectively defined due to different perceptions of space and place. One person's perception of the limits, qualities and characteristics of an urban locality will be different to others. This inherent fuzziness is unavoidable but can be given definition.

Internet Local Search tools are increasingly being used to find information regarding a place. Recent studies have shown that geographical information can be extracted by using specific search criteria (for example, Arampatzis *et al* 2004, Arampatzis *et al* 2006, Hollenstein and Purves 2006, Purves *et al* 2005, Twaroch *et al* 2009). A search may often apply to a fuzzy region such as an urban locality but the results of a search may also be partly responsible for the fuzziness of an urban locality. By linking a geographical location to internet search results an area can be given definition. Internet search results change temporally, and therefore an urban locality can be regularly redefined to reflect adjustments in internet content. Defining an urban locality has the advantage of clarifying the context of a place. Using the internet to perform this clarification takes advantage of a constantly changing resource.

Internet search results have been combined with a vector point base layer to create a tool that can assist in the definition of urban localities. Internet web pages contain geographical information, not only in the form of maps and imagery, but also within the text. For example, descriptions of an area may feature prominently in a local information website, just as a website providing a service within an area may provide an address of the service and directions for access. Such information has been used in the past to assist in the definition of areas by giving internet search results a coordinate. This approach has formed a large part of the work of the SPIRIT project (Arampatzis *et al* 2004, Arampatzis *et al* 2006, Purves *et al* 2005), as well as several other more recent research initiatives (Jones *et al* 2008, Twaroch *et al* 2009).

The aim of this work was to define the locality of Clifton in Bristol through several geographically phrased internet searches. The locality was verified with input from local estate agents, leading to questions regarding gerrymandering of boundaries for individual gain.

## 2. Methodology

### 2.1 Geo-tagging of Internet Searches

Internet searches were performed by inputting trigger phrases into the Google™ search engine. Trigger phrases are a group of words together that in this case include a reference to the location of Clifton, and place Clifton in a geographical context. An example is, "Clifton in Bristol" (see table 1 for the full Trigger phrase list). Such a phrase is used rather than just the name Clifton itself because it helps to create more geographically focused queries (Aramptzis

et al, 2006). Nine trigger phrases were input into Google (Table 1). Clifton in Bristol was found to be the most effective. This places the urban locality into the context of the Local Authority.

**Table 1.** Trigger phrases used within the Google™ search engine

Trigger Phrase 1 – "Clifton in Bristol"
Trigger Phrase 2 – "Clifton which is in Bristol"
Trigger Phrase 3 - "Clifton is a suburb in Bristol"
Trigger Phrase 4 - "Clifton is located in West Bristol"
Trigger Phrase 5 – Clifton in Bristol
Trigger Phrase 6 – "Clifton in *"
Trigger Phrase 7 – "* is located in Clifton"
Trigger Phrase 8 – "* is in Clifton"
Trigger Phrase 9 – "Clifton is close to *"

The first 100 snippets from the Google™ search Clifton in Bristol were copied into a basic text document. Snippet text is used rather than full text due to ease of retrieval and manageable handling (figure 1).

**Clifton** Online - The local portal for **Clifton Bristol** - BS8 ...
Clifton Online is a website dedicated to Clifton, Hotwells and Whiteladies Road in Bristol, including a business directory, students welcome pack, jobs, ...
www.cliftononline.net/ - 10k - Cached - Similar pages

**Figure 1.** An example Google™ search result with the snippet text highlighted yellow.

The resulting text document was searched to find post codes within. This search was performed within GATE software version 4.0 (General Architecture for Textual Engineering). The ANNIE component (A Nearly New Information Extraction System) was loaded and used to match the Google™ search results with Postcodes.

The resulting annotated text was then exported from the GATE software giving a list of Postcodes. 93% of the Postcode annotations were in the BS Postcode area which covers the Bristol region. The list was geo-tagged using coordinate information from the Ordnance Survey Code-Point® data set.

### 2.1 Composite Index of Urban Locality

A basic kernel density estimation (KDE) has been applied to the postcode points. In order to give a surface of density estimates, all postcodes within the study area (Bristol Local Authority) have been included, even where no instances of annotation were recorded (Thurstain-Goodwin and Unwin 2000). The KDE has the advantage of assigning a point density to any location inside the study region, not just to locations where there is an event (O'Sullivan and Unwin 2003, Twaroch *et al* 2009). The density is estimated by counting the number of events in a region (or kernel) centred at the location where the estimate is to be made. An intensity estimate at point p is given in

$$\hat{\lambda}p = \frac{no.[S \in C(p,r)]}{\pi r^2} \quad (1)$$

where C(p,r) is a circle of radius r centred at the location of interest p.

Figure 2 depicts the results of the kernel density estimation. The darker areas depict higher point density and a greater likelihood that a place is in the urban locality of Clifton.

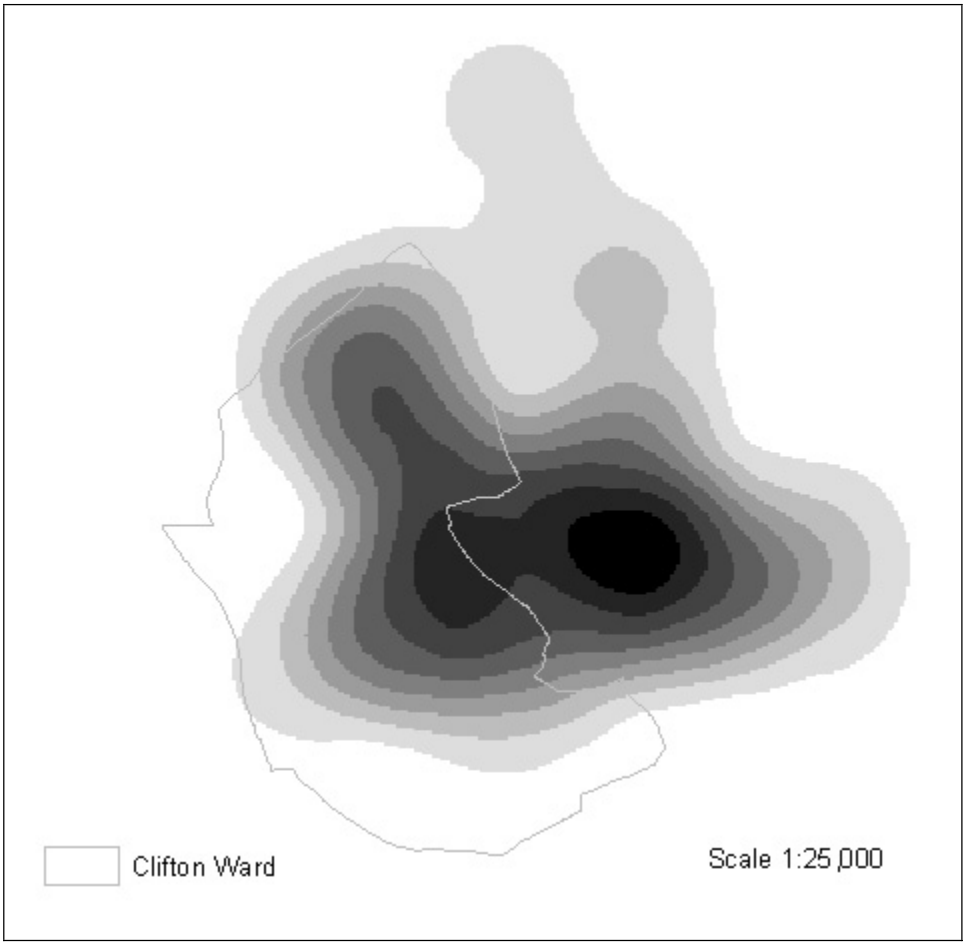



**Figure 2.** Kernel Density Estimation (KDE) of Annotated Postcode points within Bristol Local Authority

### 3. Verification of Index of Urban Locality

The results of the Index of Urban Locality were independently verified by asking eight estate agent offices to define their perception of Clifton. A similar verification technique was used in part of the SPIRIT project for verification of continuous surfaces (Purves et al, 2005). Figure 3 illustrates the Index of Urban Locality with the eight responses overlaid.

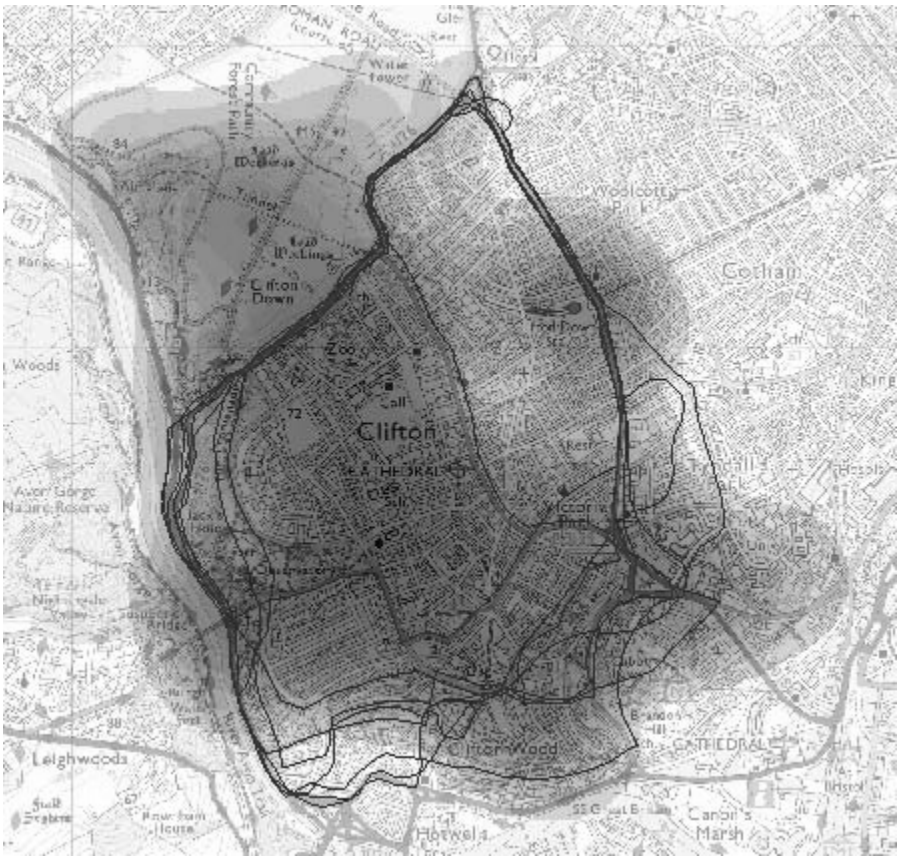





**Figure 3.** The urban locality displayed at 1:25,000 scale with respondents' interpretation of the boundary of Clifton overlaid.

It can be seen that respondents' interpretation of Clifton is consistent in the north and west. All respondents define the boundary of Clifton to the north in the same location, the A1476 road, and no respondent has defined the boundary of Clifton to be further west than the Avon Gorge.

However, definition of the boundary of Clifton is inconsistent in the south east of the region. The Index of Urban Locality is therefore useful for giving definition to Clifton, particularly in this region.

## 4. Transferability

Postcodes from the first 100 matches from the internet search Bedminster in Bristol have been annotated and geo-tagged. A kernel density estimation (Equation 1) has then been applied (figure 4).

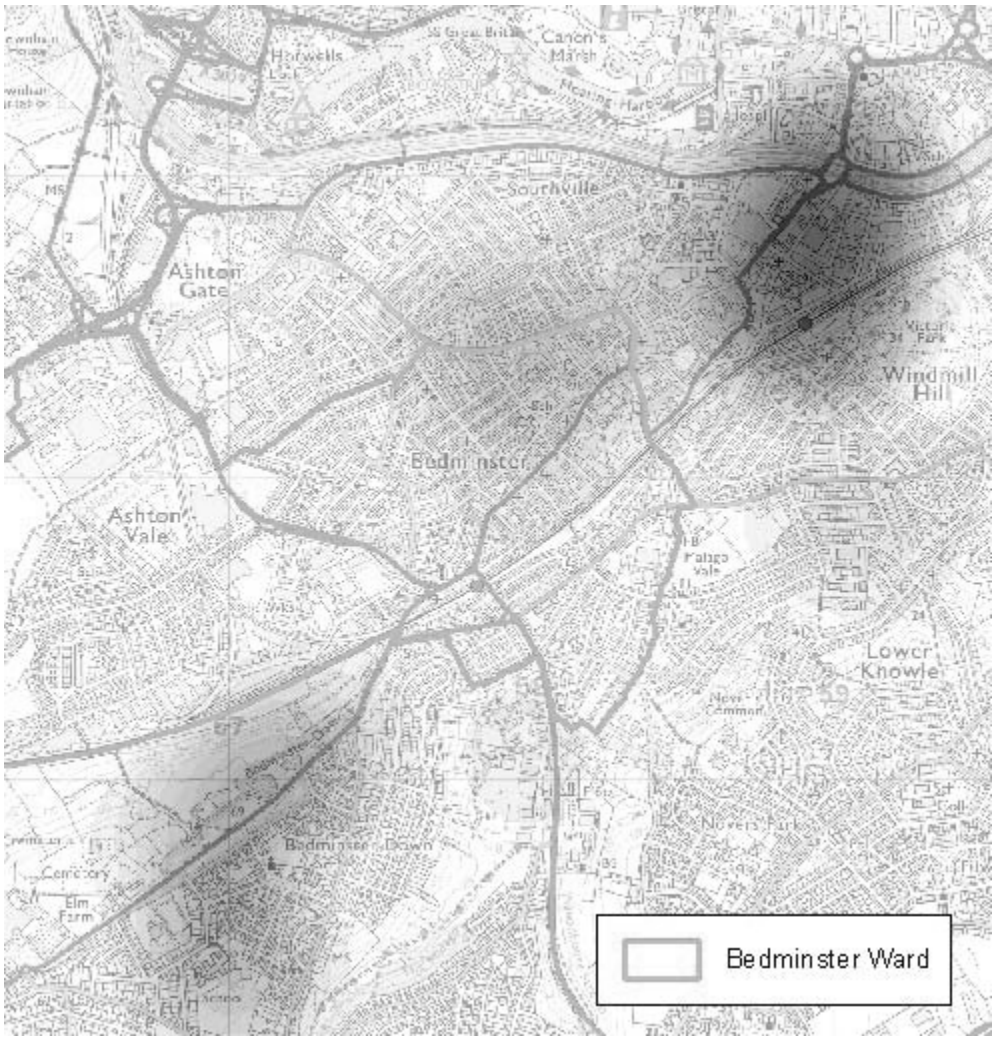





**Figure 4.** The Urban Locality of Bedminster, Bristol

It is evident from figure 4 that the urban locality of Bedminster is most prominent outside of Bedminster Ward. The darkest area can be seen to the north east. It is evident in figure 4 that when overlaid on the Ordnance Survey raster, the area that shows the strongest urban locality is large buildings of commercial land use. This may be caused by a high number of websites for businesses in this area.

The resulting Index of Urban Locality for Bedminster demonstrates that it is possible to transfer this tool to other regions. The tool depicts the urban locality of Bedminster to be different to the Bedminster Ward, and illustrates its fuzzy boundary. However, caution must be taken when making judgements regarding an area such as Bedminster without any further form of reference for verification purposes.

## 5. Conclusion

The internet is suitable as a search tool for defining the urban locality of Clifton. A methodology has been developed that searches the internet using the trigger phrase [Urban Locality] in [Local Authority]. The output is a surface visualisation that portrays membership strength. Due to the constantly changing nature of the input source (the internet), the index can be updated regularly.

It has been recognised that the methods used here are subject to the uncertainty introduced by the modifiable areal unit problem. Such recognisable uncertainties are tolerable in order to begin to overcome the vagueness that is inherent in an urban locality.

## 6. Acknowledgements

I would like to thank Vlasios Voudouris for his initial assistance in the proposal stages and his continued support throughout the study, Daniel Murtagh for his assistance in proof reading, and all estate agent employees who took the time to undergo an informal interview as part of the study.
Thanks also to the Ordnance Survey Research Department for instigation of the initial idea for this study, and for the supply of data to ensure its successful completion.

## References

Arampatzis, A., van Kreveld, M., Reinbacher I., Jones, C.B., Vaid S., Clough, P. Joho, H., Sanderson, M. (2006) Web-based delineation of imprecise regions *Computers, Environment and Urban Systems*, 30, pp436-45

Arampatzis, A., van Kreveld, M., Reinbacher I., Jones, C.B., Vaid S., Clough, P. Joho, H., Sanderson, M., Benkert,B M. and Wolff A. (2004) Web-Based Delineation of Imprecise Regions *In Proceedings of Workshop on Geographic Information Retrieval, SIGIR 2004, Sheffield, UK*.

Chang K. T. (2006) *Introduction to Geographic Information Systems* Mcgraw Hill.

Cross, V. and Firat, A. (2000) Fuzzy objects for geographical information systems *Fuzzy Sets and Systems* 113 pp19-36

Cunningham, H. et al (2007) Developing Language Processing Components with GATE Version 4 (a User Guide) The University of Sheffield http://gate.ac.uk/sale/tao/

Goodwin, M. and Unwin, D. (2000) Defining and delineating the Central Areas of Towns for Statistical Monitoring Using Continuous Surface Representations *Transactions in GIS* 4(4) pp205 -317

Hollenstein, L. and Purves, R. (submitted) Exploring place through user-generated content: using Flickr to describe city cores *Journal of Spatial Information Science*

Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M. J., & Weibel, R. (2002) Spatial information retrieval and geographical ontologies—an overview of the spirit project *In Proceedings of the 25th annual international conference on research and development in information retrieval (SIGIR 2002)* pp387–388

Jones, C. B., Purves, R., Clough, P. and Joho, H. (2008) Modelling Vague Places with Knowledge from the Web *International Journal of Geographical Information Systems* 22(10) pp1045-1065

O'Sullivan, D. and Unwin, D. J. (2003) *Geographic Information Analysis*. Wiley

Purves, R., Clough, P., Joho, H., Jones, C., Van Kreveld, M. (2005) Modelling Vague Places with Knowledge from the Web SPIRIT project publication D24 3301

Purves, R., Clough, P. D., & Joho, H. (2005) Identifying imprecise regions for geographic information retrieval using the Web *In Proceedings of GISRUK 2005* pp. 313–318

Twaroch, F. A., Jones, C. B., Abdelmoty, A. I. (2009) Acquisition of Vernacular Place Names from Web Sources *In Weaving Services and People on the Worldwide Web* pp 195-214

Twaroch, F. A., Purves, R. S., Jones, C. B. (2009) Stability of Qualitative Spatial Relations between Vernacular Regions Mined from Web Data *In Proceedings of Workshop on Geographic Information on the Internet, Toulouse, France, April 2009*

Twaroch, F. A., and Jones, C. B. (2010) A Web Platform for the Evaluation of Vernacular Place Names in Automatically Constructed Gazetteers *In Proceedings of Workshop on Geographic Information Retrieval, Zurich, Switzerland, 2010*

Wang, F., and Hall, B. (1996) Fuzzy Representation of Geographical Boundaries in GIS *International Journal of Geographical Information Systems* 10(5) pp573-590

Worboys M. and Duckham M. (2004) *GIS A Computing Perspective*. CRC Press.

**Biography**

*I am a Technical Manager for Blom Aerofilms Ltd and have an in depth knowledge of remote sensing acquisition techniques, spatial data handling practices, and geographic visualisation. I am currently undertaking a part-time PHd specialising in vernacular geography and information retrieval at Birkbeck College, University of London.*

# Topological Consistent Generalization of OpenStreetMap

Padraig Corcoran[1], Peter Mooney[1], Adam Winstanley[1]

[1]National University of Ireland Maynooth,
Maynooth, Co. Kildare,
Ireland.
Tel. +353-1-708 3847
padraigc@cs.nuim.ie

KEYWORDS: OpenStreetMap, Topological Consistent, Generalization.

## 1. Introduction

Since its introduction in 2004, OpenStreetMap (OSM) has become an important source of geospatial information. The spatial data displayed by OSM represents the finest level of detail available. For web-based and mobile mapping applications it is often desirable to reduce the representation of such spatial data using generalization. There are two main reasons for this. Firstly, the spatial size and detail of OSM data is ever increasing and the bandwidth required to transmit this can be significant. Any device which attempts to obtain such data over a network will have finite bandwidth and this may prevent the map being transmitted in its original form. Secondly, many mobile devices used to display OSM data have limited screen resolution and processing power, and can only display a finite amount of detail. One class of generalization algorithm known as simplification methods, attempt to generalize polygon and line features by reducing the number of points or vertices used to represent them. Any simplification technique can be classified as a decomposition or reconstruction technique. Decomposition techniques produce a generalized result by iteratively removing points from the original polygon or line feature in question until a desired scale is reached. Reconstruction techniques produce a generalized result by initially representing the feature in its simplest form; subsequently, points are added in an iterative manner until a desired scale is achieved.

The intention of map generalization is to produce the best result possible subject to a set of objectives (Jones and Ware 2005). Topological objectives are primarily concerned with the need to ensure that the simplified representations of the selected features retain original relationships of containment and connectivity. The research presented in this work focuses on strategies for determining if a given simplification satisfies topological objectives; that is, it is topologically consistent. Any method for determining topological consistency may be summarised in terms of the following constraints:

- Constraints on the types of topology for which the technique can determine consistency without returning a false-positive; that is, classifying a simplification as topologically correct when in fact it is not.
- Constraints on the types of topology for which the technique can determine consistency without returning a false-negative; that is classifying a simplification as topologically incorrect when it is in fact correct.
- Constraints on the types of simplification to which the technique can be applied.

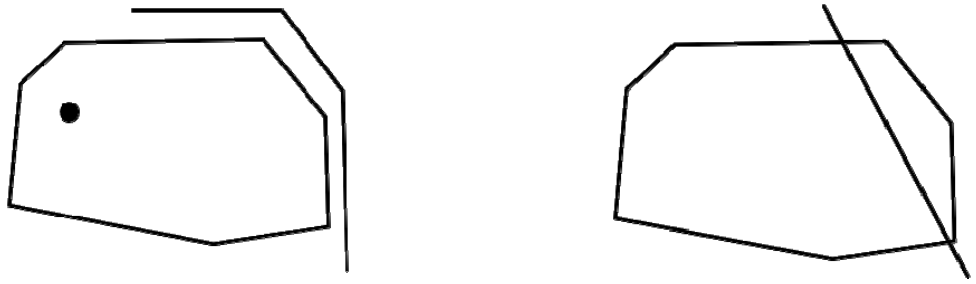
If a technique has no such constraints it may be regarded as optimal. Developing a technique which is optimal is this sense is the research contribution made by this work. To achieve this goal we performed a geometrical analysis of strategies for determining the topological consistency of a vector map simplification. We propose that all topological relationships may be classified as planar or non-planar. A formal analysis of techniques for determining consistency in terms of such relationships was performed. For each technique we analysed any corresponding constraints which are imposed. This provides a unified understanding of the benefits and limitations of individual techniques and the relationships which exist between techniques. A new approach for determining non-planar

topological consistency, which imposes the least possible constraints, is proposed. The effectiveness of this approach is demonstrated through the fusion with an existing simplification technique.

The layout of this paper is as follows. In the following section we introduce the concepts of planar and non-planar topological features. In section 3 we briefly review and summarize all existing techniques for determining planar and non-planar topological consistency. Through this analysis the authors developed a new strategy for determining topological consistency. In section 4 we present results and draw conclusions.

## 2. Planar and non-planar topological properties

All topological properties can be classified as those which represent a planar embedding of a graph and those which do not. Consider the simple map in Figure 1(a) which contains a polygon, line and point feature. No lines or edges in this map cross without forming a vertex; therefore we say that the topological relationships between all features are planar. Next consider the simple map in Figure 1(b) which contains a single polygon and line feature. Due to the fact that the line crosses the polygon without forming a vertex, we say the topological relationship between these features is non-planar.



(a) (b)

**Figure 1**. Planar relationships are displayed in (a). Non-planar topological relationships are shown in (b).

## 3. Maintaining topological consistency

We analysed three seminal methods for determining planar topological consistency in terms of the relationships for which they can determine consistency and constraints they impose. These are the works of De Berg et al. (1998), Saalfeld (1999) and da Silva and Wu (2006). Table 1 represents a summary the results of this analysis. For each technique we indicate if it can determine if a given simplification is topological inconsistency with respect to the three types of planar topological relationships. An asterisk indicates that the technique in question can determine the topological consistency without constraint and is therefore optimal. An $x_i$ symbol indicates that the technique in question can determine the topological consistency but is subject to some form of constraint and therefore is not optimal. The final column defines what each constraint $x_i$ represents. From this table the following conclusions can be drawn. The consistency of point features should be determined by the strategy of De Berg et al. (1998) while the consistency of line features should be determined by the strategy of da Silva and Wu (2006). This strategy imposes the least possible constraints.

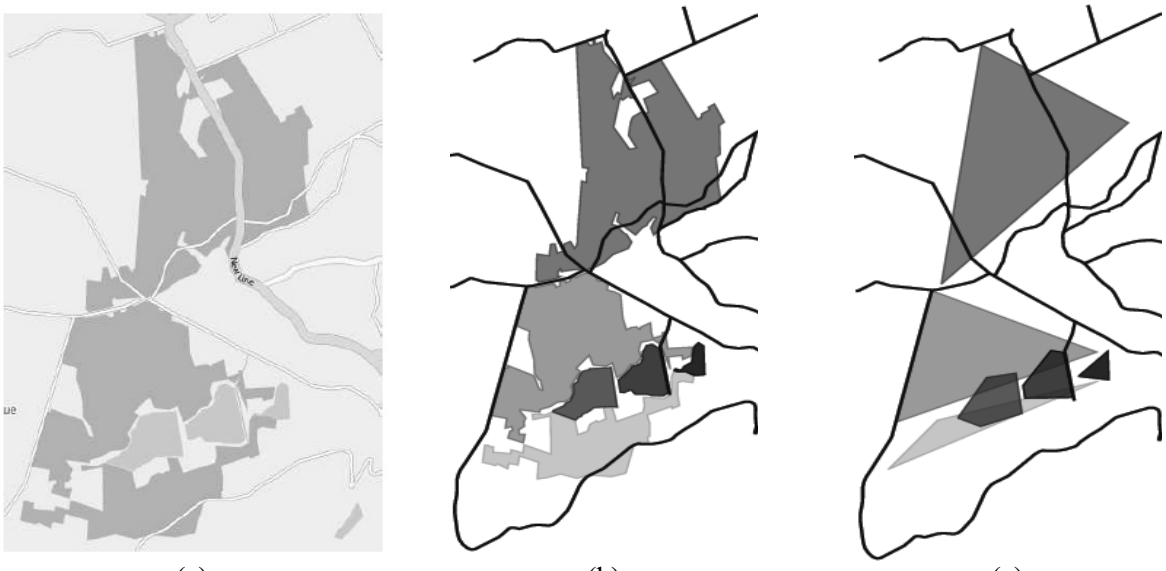
Most existing works proposed to maintain the non-planar topology line intersections by simply maintaining all line segments that contain intersections in the map. The drawback of this strategy is that it severely restricts the number of possible simplifications for which consistency can be determined. In this work we developed a technique which overcomes this limitation.

**Table 1**. The forms of planar topology for which each technique can determine consistency and any corresponding constraints.

| | Point | Line | Non self-intersection | Constraints |
|---|---|---|---|---|
| de Berg | $x_1$ | | $x_1$ | $x_1$ – Monotone chains |
| Saalfeld | * | | | |
| da Silva | $x_2$ | * | * | $x_2$ – On point features |

## 4. Results and Conclusions

In section 3 methodologies which can determine planar and non-planar topological consistency of a given simplification were presented. We propose to fuse these with an existing simplification technique which satisfies shape objectives. To satisfy shape objectives a contour evolution or simplification technique proposed by Latecki and Lakmper (1999) was used. Consider the sample map taken from OSM which is displayed in Figure 2(a) and plotted in Figure 2(b). The polygons in this map represent forests and lakes; the line features represent a road network. Each polygon in this map was simplified using the contour evolution technique with the corresponding result shown in Figure 2(c). Although the simplification satisfies shape objectives, it contains many topological inconsistencies. This includes the introduction of overlapping polygons and the removal of polygon line intersections.



(a) (b) (c)

**Figure 2**. The map in (a) is plotted in (b) and simplified in (c).

To demonstrate the effectiveness of proposed topological consistent simplification technique, the polygons in this data set were simplified by an increasing amount. These results are shown in Figure 3. The final simplification represents the convergence of the algorithm. From these results we can see that all planar topological relationships mentioned above were maintained throughout each simplification scale. It is also evident that each simplification satisfies shape objectives and represents a progressive intuitive shape evolution. The number of vertices used to represent each polygon at each simplification stage is given in Table 2. The final simplification in Figure 3(f) represents a 91% reduction in the number of vertices relative to the original map.

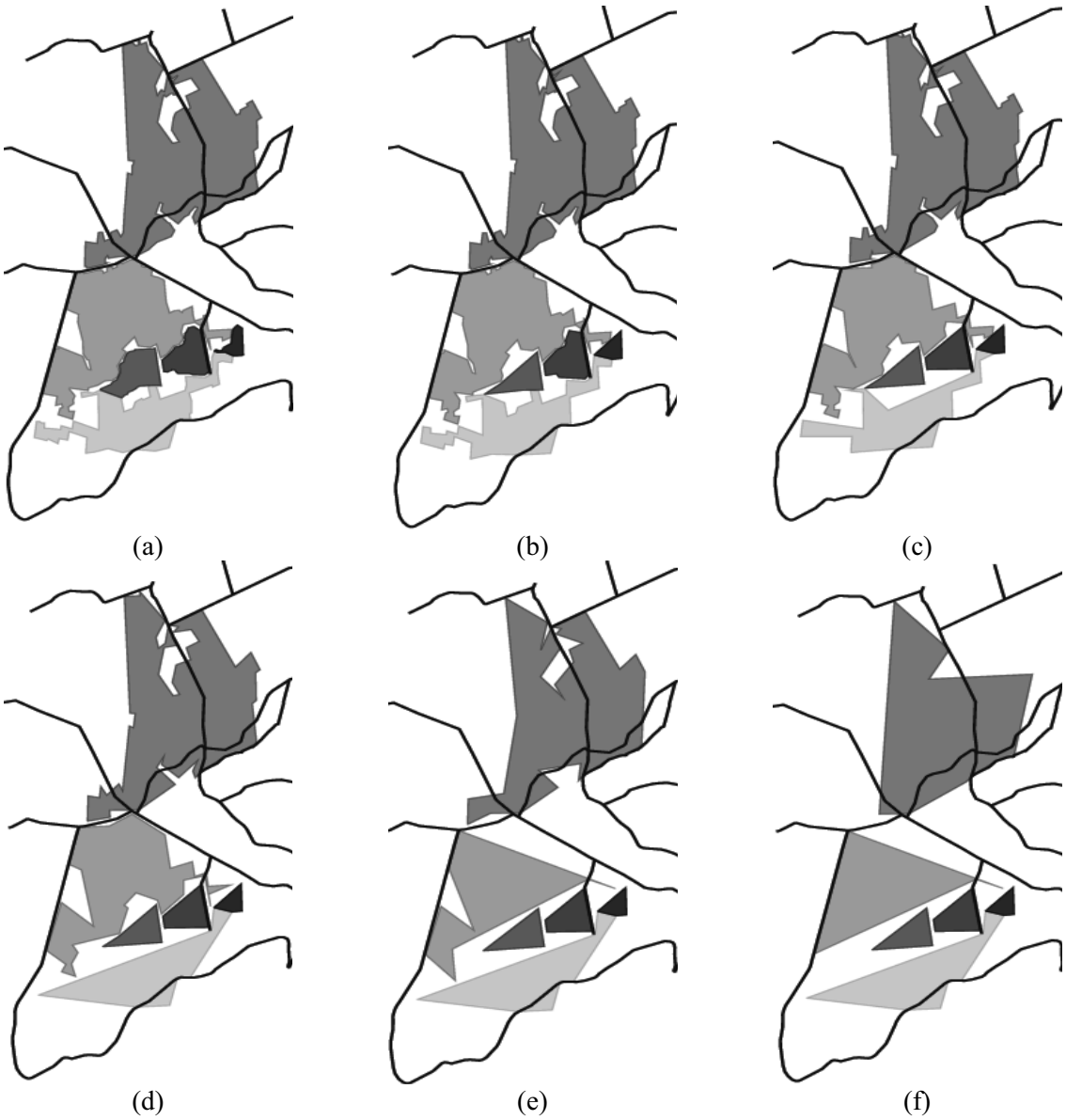



**Figure 3**. The set of polygon in (a) are simplified by an increasing amount in (b)-(f).

**Table 2**. The number of polygon vertices at each simplification stage in Figure 3.

| | Poly. 1 | Poly. 2 | Poly. 3 | Poly. 4 | Poly. 5 | Poly. 6 | Total | % Red. |
|---|---|---|---|---|---|---|---|---|
| Fig. 3(a) | 20 | 35 | 28 | 126 | 101 | 65 | 375 | 0 |
| Fig. 3(b) | 4 | 12 | 5 | 102 | 78 | 42 | 243 | 35 |
| Fig. 3(c) | 4 | 4 | 4 | 79 | 55 | 16 | 165 | 56 |
| Fig. 3(d) | 4 | 4 | 4 | 56 | 32 | 8 | 108 | 71 |
| Fig. 3(e) | 4 | 4 | 3 | 33 | 9 | 8 | 61 | 84 |
| Fig. 3(f) | 4 | 4 | 3 | 11 | 4 | 8 | 34 | 91 |

Due to space constraints, this paper only represents a very brief description of how our proposed generalized algorithm maintains planar and non-planar topological consistency. A more in-depth analysis will be presented in a later publication. Although this work has focused on the task of simplifying polygon features, the proposed methodology could also be applied in the context of simplifying line features. This will be the focus of future research.

**6. Acknowledgements**

Research presented in this paper was part-funded by a Strategic Research Cluster grant (07/SRC/I1168) from Science Foundation Ireland under the National Development Plan.

**References**

Yang B, Purves R and Weibel R (2007) Efficient transmission of vector data over the Internet *International Journal of Geographical Information Science* **21(2)** pp215-237

Latecki L and Lakmper R (1999) Convexity rule for shape decomposition based on discrete contour evolution *Computer Vision and Image Understanding* **73(3)** pp441 – 454

Saalfeld A (1999) Topologically Consistent Line Simplification with the Douglas-Peucker Algorithm Cartography and Geographic Information Science **26(1)** pp7-18

De Berg M, van Kreveld M, and Schirra S (1998) Topologically correct subdivision simplification using the bandwidth criterion *Cartography and Geographic Information Systems* **25(4)** pp243–257.

da Silva A and Wu S (2006) A Robust Strategy for Handling Linear Features in Topologically Consistent Polyline Simplification *GeoInfo* pp19–34.

**Biography**

*Padraig Corcoran completed a B.Sc. in Computer Science and Software Engineering and a Ph.D. in Computer Science in 2004 and 2008 respectively. He is currently working as a lecturer and researcher in the department of Computer Science at the National University of Ireland Maynooth.*

*Peter Mooney holds a BSc (1999) and PhD (2004) in Computer Science from the National University of Ireland Maynooth (NUIM). He is currently a Postdoctoral Research Fellow with the Department of Computer Science at NUIM and at the Irish Environmental Protection Agency (EPA).*

*Adam Winstanley gained MSc and PhD degrees in Computer Science in 1987 and 1991 respectively. Currently, he is Senior Lecturer and Head of Department of Computer Science and Senior Research Associate of the National Centre for Geocomputation at NUI Maynooth.*

# *Matterhorn on the Horizon*: Identification of Salient Mountains for Image Annotation

Martin Tomko[1], Ross S. Purves[1]

[1]Department of Geography, University of Zurich, Switzerland
Tel. (+ 41 44 635 5256) Fax.(+ 41 44 635 6848)
martin.tomko|ross.purves@geo.uzh.ch, www.geo.uzh.ch/~mtomko

KEYWORDS: mountain salience, DEM, image annotation, toponym

## 1. Introduction

Information retrieval searches indexes to identify relevant content for a particular query. While in the case of plain or structured text documents, the index is typically generated from document text, image search is often based on assigned keywords. Adding keywords to images to allow their effective retrieval is an important element of the work of commercial image libraries, and one associated with high costs (keyword annotation is laborious) and benefits (well annotated images are retrieved by customers and thus purchased). The Tripod project (Purves et al., 2008) aims to exploit landscape photographs tagged with location information (for example coordinates and a direction) and automatically annotate such images based on spatial data and Web content related to image location.

Topographic eminences are a frequent and highly salient element of landscape photographs – for example, (Edwardes & Purves, 2007) found that *hill* was one of the most common concept terms used in the description of images in Geograph. Names of salient hills and mountains are thus important keywords in the annotation of landscape photographs. This is especially true for mountains belonging to the horizon line – they stand out against the background formed by the sky (they are therefore salient), they often have a recognisable silhouette and they provide a clear structural element in the photograph.

The question then is how to identify and name visually salient mountains visible from a specific vantage point? Note that often the summit itself need not be visible, but the mass of the mountain would still be named by a human. Our hypothesis is that terrain analysis methods alone allow for accurate identification of salient mountains.

We present a method based on a sequence of analytical steps, assigning toponyms to mountain-top regions using relative drop, computation of horizon lines based on viewshed analysis, and identification of the most salient points on the horizon based on a hierarchical decomposition of the projected horizon. In the approach presented we explicitly assume free visibility to the horizon line, in other words, we do not consider occlusion of the horizon by vegetation or man-made structures, as discussed by Tomko et *al.* (2009).

## 2. Background

Viewshed computation on digital elevation models (DEMs) is a popular terrain analysis technique, central to the identification of salient mountains. It has been researched by the GI community from many perspectives, such as algorithms and DEM data structures (De Floriani & Magillo, 2003), accuracy and veracity of viewsheds (Maloy & Dean, 2001), and applications, for instance, in visual impact assessment, (Ervin & Steinitz, 2003). Fisher (1996) proposed an extension to the binary raster viewshed by classifying raster cells as either visible, invisible, on local horizon or on global horizon. We use such a classification in our image annotation approach, as detailed in Section 3.

Outdoor enthusiasts and geomorphometrists alike are interested in assessing the prominence of

mountains. Classifications derived from elevations of peaks relative to the surrounding terrain are popular for their computational simplicity and relative objectivity (Grimm & Mattmüller, 2004; Helman, 2005; Podobnikar, 2009). There is, however, no direct relationship between such derived mountain prominence and the apparent salience of an observed peak. In a given horizon line, the summit of the peak may not be directly visible, and typically, local minima are formed by perceivable overlaps of the silhouettes of closer and more distant hills and not by defining saddles. As noted by Fisher et *al.* (2004), the extent of mountains is larger than the summit, and it is hard to delineate. The visibility of the mountain massif, however, seems a sensible element to refer to in image annotation – in other words we name the region of mountain top, not merely its summit.

Chippendale et *al.* (2008) combined DEM and image content analysis to evaluate the salience of features found within artificially constructed landscape images. Their approach does not, however, provide details about the algorithms used and furthermore requires the use of, typically, computationally expensive image content analysis. As such, it is not well suited for real-time image annotation.

## 3. Method

The method proposed relies on the computation of the enriched viewshed from the image origin, consequent identification of the cells belonging to the most salient apparent peak by hierarchical decomposition of the global horizon and finally by matching the cells belonging to the salient peak with the containing mountain-top regions identified by DEM analysis using relative drop.

We use the freely available SRTM DEM 90m dataset for viewshed computation and the identification of the mountain-top regions. The SwissNames toponym database for Switzerland (containing toponyms from 1:25000 topographic maps) is used to relate mountain-top regions with the names of the mountains constituting the regions.

An enriched viewshed is computed for a given vantage point specified by WGS84 coordinates automatically extracted from the EXIF metadata of a georeferenced image. The computation can be limited to a certain distance and an oriented field of view. Points that belong to the global horizon are selected and ordered by azimuth. We store the horizon line in a vertical coordinate system defined by the azimuth and the elevation angle of the observed cell. The horizon line is smoothed to remove artefacts from the viewshed analysis and DEM data. It is then searched for local minima and maxima, and the average relative height of each of the maxima, defined as a height difference between the maximum and the surrounding two local minima is computed. The result is a hierarchical list of the maxima with the largest height difference on the horizon line, related to the apparent prominence of the peak.

The cell identified as belonging to the most salient local maximum of the horizon line is related to the peak region it belongs to. The mountain-top region dataset used contains 531 regions (polygon geometries) over the whole of Switzerland, formed by mountains over 800m in altitude with a relative drop to the nearest saddle of over 250m. The regions do not represent a complete tessellation of the area. As any given mountain-top region can be generated by multiple named peaks (e.g. auxiliary summits, such as in the massif in the right of Figure 1), we select the peak observed under an azimuth closest to that of the most salient horizon maximum. The identified peak's name is used in image annotation, together with the indication of the peak's position in the image (Image left, middle or right).

## 4. Example Output and Further Research

Figure 1, showing a typical mountain scene from the Swiss Alps, has been automatically annotated with the toponym *Chöpfenberg*, with the position of the detected mountain indicated as image left. Indeed, the notable rocky mountain that occupies the left side of the image background is Chöpfenberg.

**Figure 1.** A view of Swiss Alps, Chöpfenberg in the background on the left.

The method selects salient mountains exclusively by analysing the horizon line (Figure 2). Further optimization is possible by considering only toponyms from topographical maps of a certain scale (e.g. 1:100 000 vs. 1:25 000,) to avoid selection of subsidiary summits. Such a selection allows for de-facto cartographic generalisation in salience estimation. People are likely to consider further aspects - specific shape of mountain silhouette, texture, geological and cultural significance to name a few - when selecting mountains for annotation of images. These considerations are subject to future research.

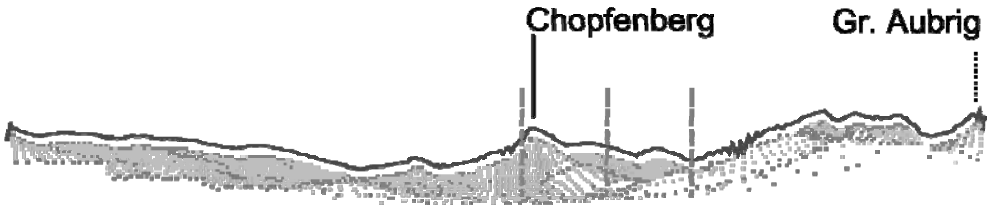



**Figure 2.** The 360$^{o}$ horizon containing the scene from Figure 1 (extent indicated by dashed red lines for margins and center), constructed from a viewshed with limit view distance of 3.8km. The location of the *Chopfenberg* indicated by the black line, a second salient summit (*Gross Aubrig*) indicated further towards the right is outside Figure 1.

**5. Acknowledgements**

This research reported is part of the project *TRIPOD* supported by the European Commission under contract 045335. We would also like to acknowledge the SwissNames dataset provided by the Federal Office of Topography Swisstopo and the SRTM 90m DEM dataset provided freely by CGIAR.

**References**

Chippendale, P., Zanin, M., & Andreatta, C. (2008). *Spatial and Temporal Attractiveness Analysis through Geo-Referenced Photo Alignment.* Paper presented at the IEEE International Geoscience & Remote Sensing Symposium, July 6-11, 2008, , Boston, Massachusetts, U.S.A.

De Floriani, L., & Magillo, P. (2003). Algorithms for Visibility Computation on Terrains: a Survey. *Environment and Planning B: Planning and Design, 30*(5), 709-728.

Edwardes, A. J., & Purves, R. (2007). A Theoretical Grounding for Semantic Descriptions of Place. In J. M. Ware & G. E. Taylor (Eds.), *Web and Wireless Geographical Information Systems. 7th International Symposium, W2GIS 2007, Cardiff, UK, 28-29th Nov. 2007, Proceedings* (Vol. 4857, pp. 106-120). Berlin, Heidelberg: Springer-Verlag.

Ervin, S., & Steinitz, C. (2003). Landscape Visibility Computation: Necessary, but not Sufficient. *Environment and Planning B: Planning and Design, 30*(5), 757-766.

Fisher, P., Wood, J., & Cheng, T. (2004). *Where is Helvellyn?* Fuzziness of Multiscale Landscape Morphometry. *Transactions of the Institute of British Geographers, 29*(1), 106-128.

Fisher, P. F. (1996). Extending the Applicability of Viewsheds in Landscape Planning. *Photogrammetric Engineering & Remote Sensing, 62*(11), 1297-1302.

Grimm, P., & Mattmüller, C. R. (2004). *Die Gebirgsgruppen der Alpen : Ansichten, Systematiken und Methoden zur Einteilung der Alpen.* (Vol. 39). Munich, Germany: Deutscher Alpenverein.

Helman, A. (2005). *The Finest Peaks*. Victoria, BC, Canada: Trafford Publishing.

Maloy, M. A., & Dean, D. J. (2001). An Accuracy Assessment of Various GIS_Based Viewshed Delineation Techniques. *Photogrammetric Engineering & Remote Sensing, 67*(11), 1293-1298.

Podobnikar, T. (2009). *Method for Determination of the Mountain Peaks.* Paper presented at the 12th AGILE International Conference on Geographic Information Science 2009, Leibniz Universität Hannover, Germany.

Purves, R., Edwardes, A. J., & Sanderson, M. (2008). Describing the Where – Improving Image Annotation and Search Through Geography. In G. Csurka (Ed.), *Proceedings of the 1st Intl. Workshop on Metadata Mining for Image Understanding (MMIU 2008)* (pp. 105-113). Funchal, Madeira – Portugal.

Tomko, M., Trautwein, F., & Purves, R. (2009). Identification of Practically Visible Spatial Objects in Natural Environments. In M. Sester, L. Bernard & V. Paelke (Eds.), *Advances in GIScience* (pp. 1 - 23). Berlin Heidelberg: Springer-Verlag.

**Biography**

*Martin Tomko is a post-doctoral research assistant at the Department of Geography of the University of Zurich, where Ross Purves is a lecturer.*

# Planning Alerts for Community Maps

Yang Liu, Claire Ellul, Muki Haklay

Dept. of Civil, Environmental and Geomatic Engineering, University College London,
Gower Street, London, WC1E 6BT
Tel. +44 (0) 20 7679 4118 Fax +44 (0) 20 7380 0453
Email : uces012, c.ellul, m.haklay@ucl.ac.uk

KEYWORDS: planning application, PPGIS, GeoWeb

## 1. Introduction

The UK government has clearly stated its ongoing commitment to involving communities and citizens in the decision-making process in a recently published government White Paper entitled "Communities in Control – Real People, Real Power" (Department of Communities and Local Government, 2008). In this paper, access to local information was identified as a prerequisite to encourage participation in local democracy. Such information enables citizens to better understand the services and facilities available in their locality and become more involved in local decision-making.

Building on this concept, the Mapping Change for Sustainable Communities (MCSC) project and its follow-on social enterprise Mapping for Change make use of web mapping technology to help communities better understand and become involved in changes occurring in their local area (Ellul *et al*. 2008). A Community Map has been created for each participant group to allow them to capture and maintain information of their choice about their local environment and to act as a one-stop-shop for local information. Themes identified by the local community groups range from local shops and services to recycling facilities, sources of noise and pollution and historical information about the neighbourhood. Additionally, planning and development information has been identified as important by all groups. Maps can be found at http://communitymaps.org.uk.

This paper presents an enhancement to the original Community Maps website to enable groups to become more involved in the planning application decision-making process by providing them with timely access to submitted planning applications. In particular, the work described utilises a technique to automatically incorporate third-party information on the maps which contrasts with the manual data capture processes previously required. Such an approach greatly enhances the potential of the site to act as a one-stop-shop for neighbourhood information.

## 2. Notification Procedure in the UK Planning Application System

Planning permission is a required consent when new buildings or major changes are planned to take place on existing buildings or in the local environment (Town and Country Planning Order 1995). It is normally granted by the Local Planning Authority (LPA) which, as part of the decision-making process, is required to notify all the "adjoining owners or occupiers" of the application site and take their representations into account. This notification and representation procedure is illustrated in Figure 1. Notifications can be given by site display for no less than 21 days, by advertisement in a newspaper circulating in the locality, or through an individual notification letter. Anyone can comment on a planning application.

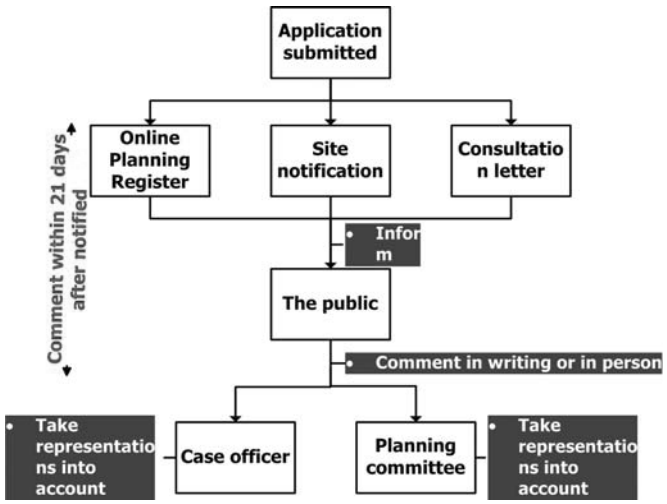



**Figure 1.** General Procedure of Planning Application Notification

The definition of notification recipients and the information dissemination means employed in the current planning application system raises doubts as to whether everyone potentially impacted by a proposed development will actually be notified. Firstly, there is no statutory obligation for the LPAs to notify citizens living in a property that does not adjoin the application site. Secondly, even residents of adjoining sites may not come across the site notice or local advertisement regarding a planning application and will therefore be denied the opportunity to make a representation. Consequently, other notification approaches are required to ensure that potentially interested parties are fully involved in the planning process.

### 3. Planning Alerts and Community Maps

Launched at the end of 2006, PlanningAlerts.com hosts planning application data sourced from the official websites of 321 local authorities out of 468 in the UK. In the Greater London area, Planning Alerts covers 32 local authorities out of 33 (Planning Alerts, 2009). Planning Alerts allows users to sign up for email alerts detailing planning applications in their neighbourhood, and also provides access to the hosted planning application data via Application Programming Interface (API).

Planning Alerts data is provided in the form of a geographic Really Simple Syndication (geoRSS) feed, which automatically publishes new planning data, incorporating coordinate information, on a daily basis. The published data, which is presented as an XML document, is then automatically collected and parsed by the Community Maps code. Each application is inserted in the Community Maps database and then displayed on the relevant Community Map, which shows planning application records published within the previous 10 days. Associated information includes the address of the proposed site, the description of the development, the publication date of the application, the unique ID of the application, the link to the application record and the link to make comment on the relevant LPA's website. Users can then access planning application information by visiting the Community Maps website.

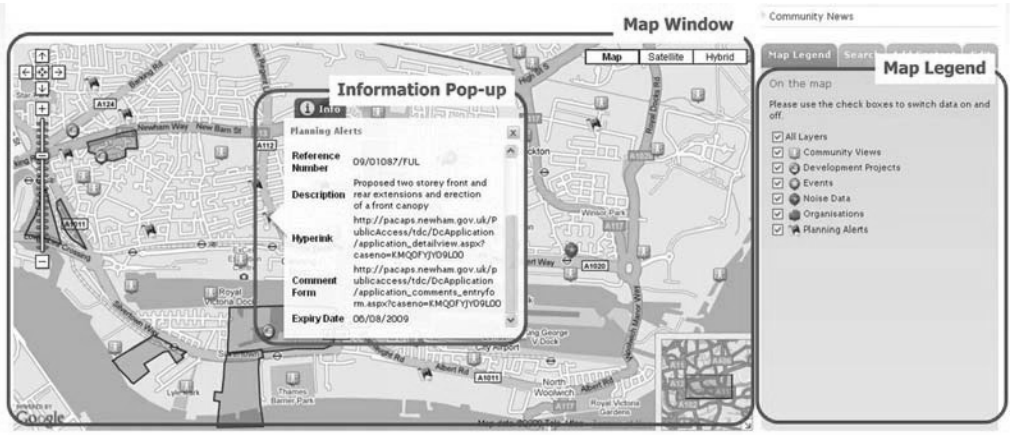



**Figure 2.** Planning Alerts on Community Maps

**4. Bridging the Digital Divide - Text Messaging Services for Planning Alerts**

The automated provision of planning information in a geographical context overcomes a number of the limitations of the current notification process, in particular widening timely access to planning applications through a one-stop-shop approach. However, a number of assumptions are made when utilising Web GIS for information dissemination. Firstly, it is assumed that potential users have the broadband internet bandwidth required to access the full functionality provided by such applications. However, the assumption is challenged by a figure released by the Office for National Statistics (2009) where only 63% of the UK population had broadband internet connection in 2009. Additionally, users are assumed to have the web browsing skills to operate such systems, which move beyond standard internet interaction paradigms required for searching and clicking links. Citizens with no digital access are referred to as "digitally excluded" or "digitally marginalised" (Ellul *et al*, 2008).

A survey carried out in 2007 reveals that the proportion of households who own a mobile phone was 78% (Office for National Statistics, 2008). Text message-based dissemination of information can therefore enhance the potential audience to people who do not have access to internet but own a mobile phone. As part of the Community Maps development programme, Ellul and her colleagues (Ellul *et al.* 2008) developed the EcoTEXT system which sends custom text messages to subscribers in accordance with their registered preferences. To date, content of the messages relates to neighbourhood events, with event information uploaded manually by users.

This service has now been extended to Planning Alerts data, and users wishing to subscribe to Planning Notifications can now register their preferences through Community Maps or by telephone. These preferences include the postcode of their residence, the distance from their residence to the farthest proposed development site that they want to be notified of, the maximum number of messages they would like to receive per week and their mobile number (Figure 3). Community Maps will then automatically send timely planning application information matching the selected preferences, querying the incoming stream of planning application data on a daily basis. Alerts can also be sent via E-mail to those users who do have internet access but do not have broadband and/or sufficient skills to use a web-mapping application.


**Figure 3.** User Preferences Registration Form

A third-party SMS gateway service provided by Sponge Ltd (Sponge Ltd, 2009) is used to dispatch messages to their intended recipients. Although messages are limited to 160 characters, each message contains sufficient information to allow users make an informed decision about potential representations (the address and brief description of the proposed plans). Examples of the resulting messages are shown in Figure 4 , with a ">" character used to separate each element.

**From:** CommuniMaps

**Message Content:** 10 Spurgeon Road, Upper Norwood, SE19 3UF> Demolition of garage and bomb shelter; erection of two bedroom detached chalet bungalow

**From:** CommuniMaps

**Message Content:** Macmillan House Paddington Station Praed Street W2 1BA> Provide two integrated office spaces by creating doorways in adjoining office walls between office B115 and B116.

**Figure 4.** Sample Messages Generated by Community Maps

## 6. Discussion and Future Work

The automated incorporation of data onto the Community Maps website contrasts markedly with the previously existing data capture process, which required manual data collection. Additionally, it opens up the potential to incorporate additional datasets of interest as they become available. The use of text messaging has the potential to deliver notification to a wider group of local citizens, in particular those who may be digitally marginalised.

Two issues remain to be resolved. Firstly, the funding model for the messaging service requires further consideration. Should this be state funded, sponsored or subscriber funded? Secondly, at the time of writing, the PlanningAlerts.com has temporarily ceased publishing planning data due to issues with licensing affecting its postcode data provider service, ErnestMarple.com (Ernest Marples Blog, 2009). This service has been taken offline due to legal action taken by Royal Mail for "improperly using their postcode database". It is hoped that the Planning Alerts service will start again in the near future (in particular due to the wider free availability of UK data due in April 2010, BBC 2009). Further work will then concentrate on deploying the mapping and text messaging services in a live environment so that the actual impact can be evaluated.

## References

BBC (2009), *Ordnance Survey Maps to go Online*, [Internet], Available from: http://news.bbc.co.uk/1/hi/technology/8366190.stm, [Accessed 22nd November 2009]

Department of Communities and Local Government. (2008), *Communities in Control, Real people, real power – Government White Paper* [Internet], Available from: http://www.communities.gov.uk/communities/communityempowerment/communitiesincontrol/ [Accessed August 9 2009]

Ellul, C., Haklay, M. and Francis, L. (2008) Empowering individuals and community groups – is Web GIS the way forward? In: AGI *GeoCommunity '08: Shaping a Changing World*, September 23 - 25, 2008, Stratford upon Avon, UK.

Ernest Marples Blog (2009) The blog for Ernest Marples' Postcodes [Internet], Available from: http://ernestmarples.com/blog/ [Accessed October12 2009]

Office for National Statistics (2008) *Consumer Durables* [Internet] Available from: http://www.statistics.gov.uk/CCI/nugget.asp?id=868 [Accessed August 16 2009]

Office for National Statistics (2009) Internet Access [Internet] Available from: http://www.statistics.gov.uk/cci/nugget.asp?ID=8 [Accessed Oct 20 2009]

*Planning Alerts Websites.* [Internet], Available from: http://www.planningalerts.com/ [Accessed October 10 2009]

Sponge Ltd (2009) Sponge – from Concept to Handset [Internet] Available from: http://spongegroup.com/about/, [Accessed Oct 12 2009]

*Town and Country Planning (General Development Procedure) Order 1995.* (1995) SI 1995/419 [Internet], Available from: http://www.opsi.gov.uk/si/si1995/Uksi_19950419_en_1.htm, [Accessed September 23 2009]

## Biography

*Yang Liu is a recent M.Sc. GIS graduate from the Department of Civil, Environmental and Geomatic Engineering, University College London (UCL). Her primary research interest is to help the public better utilize Volunteered Geographic Information (VGI) by developing data quality control strategies for VGI.*

# Investigating changes in the predicted probability of voter turnout when re-siting polling stations in three elections: a case study in Brent, UK

Scott Orford[1]

[1]Wales Institute of Social and Economic Research, Data and Methods (WISERD),
Cardiff University, 46, Park Place, Cardiff, Wales CF10 3WA
Tel. +44 (0)2920 875272
Email: orfords@cardiff.ac.uk | http://www.wiserd.ac.uk

KEYWORDS: voter turnout, binomial multi-level modelling, network analysis, polling stations

## 1. Introduction

Recent initiatives for increasing participation in elections have yet to replace the traditional method of voting in person at designated polling stations (Electoral Commission, 2007a). Rather, UK government policy on voting at polling stations has strengthened as has the debates about increasing voter turnout (Electoral Commission, 2007b). Recent research in the UK (e.g. Orford et al., 2009; Orford and Schuman 2002) and the US (e.g. Haspel and Knotts, 2005; Gimpel and Schuknecht 2003), has demonstrated that voters are sensitive to geographical factors, such as distance travelled to vote, the public transport network and the location of activities such as places of work and study. This is particularly true in lower salience elections such as the European parliament elections in the UK. As a consequence the choice and siting of polling stations becomes important and government policy has stated that accessibility should be a key criterion. Thus the aim of this research is to understand more fully the impact of polling station location on voter turnout in the UK by predicting the changes in the probability of voter turnout caused by re-siting polling stations in different parts of the polling district.

## 2. Siting polling stations

Although the placement of polling stations can be regarded as a simple bureaucratic detail, locating sufficient and appropriate venues for election activities is not always straightforward. In the UK it is the responsibility of local councils to designate polling stations within their area. Here, there is no legal definition of a polling station but, at a minimum, it is the room in which voting takes place (Electoral Commission, 2007b; p. 42). Traditional venues include schools, libraries and community centres but premises such as fish and chip shops, pubs, hotels and supermarkets have also been used. The UK's Electoral Commission provides guidance on polling station location and this is increasingly emphasising the need for good access. This includes both physical access to the polling station - for people with disabilities, say - and also general accessibility: "if possible, it needs to be close to where voters live and be fully accessible" (ibid. p.25). In some instances, no suitable venues are available in accessible locations and portable polling stations have to be used. Further, following the commencement of section 16 of the Electoral Administrations Act 2007, a review of polling districts and polling stations must take place at least once every four years (Electoral Commission, 2007c). Given the increased emphasis in local authority guidance on improving accessibility and reducing distances travelled to vote, an understanding of how turnout changes by siting polling stations in different locations could be part of this review.

## 3. Case study and methodology

The focus for this case study is the London Borough of Brent, which has compiled complete polling district level data on turnout for the 2001 parliamentary election, the 1999 European election and the 1998 local election. These data includes the number of voters, the number of registered electorate and the number of postal voters for each polling district. It does not contain information on the non-

registration of voters or how voters travelled to the polling station to vote (e.g. walking, by car etc.). Brent has three entire parliamentary constituencies and 31 electoral wards that are subdivided into 115 polling districts and each has a polling station where people go to vote. Electors can only cast their vote at a specified polling station. Digital boundary data and the polling station locations were supplied by Brent. The locations of the electorate within the polling districts were captured using postcode data obtained from the National Statistics Postcode Directory and each postcode was assigned to a polling district. GIS was used to construct a measure of voter density around each polling station. This is a spatial measure that captures both the aggregate distance travelled by voters to their designated polling station and their geographic distribution within the polling district. This was shown in previous research (Orford, et al. 2009) to be a superior measure of the cost of voting in person than a single average aggregate distance measure. It is also a measure of the compactness of the polling district, a characteristic that is often used to explain variations in voting behaviour in the electoral studies literature (Niemi et al., 1990). The voter density measure assumes that each voter travels from their home to the polling station (although this journey may include other activities, such as shopping or commuting) using network distances calculated from individual postcodes to the elector's assigned polling station.

Voter density was captured by creating a series of variables that measure the number of domestic addresses per postcode within specific network distances from the polling station. These specific network distances were measured at 100 metre intervals from the polling station up to a maximum of two kilometres (the furthest network distance between a postcode and a polling station in Brent). This created twenty voter density variables (one for each interval), each capturing the cumulative numbers of domestic addresses per postcode with increasing distances from the polling station. These numbers were then converted into proportions representing the share all domestic addresses in the polling district. Hence the first voter density variable captures the proportion of domestic addresses within 100 network distance metres of the polling station as a share of all domestic addresses in the polling district; the second variable captures the proportion of domestic addresses within 200 network distance metres of the polling station as a share of all domestic addresses in the polling district and so on.

Since the objective of the research is to investigate the impact of re-siting the polling station to new locations within a polling district, the voter density of each postcode in Brent was also calculated. This was a substantial undertaking with the procedure described above applied to all 8841 postcodes in Brent. Then, for each polling district, the postcodes that represented that maximum, minimum and average voter density location were identified. By relaxing the constraint that polling stations tend to be located within particular buildings, all postcode locations can be investigated rather than simply those with a suitable building. Figures 1 & 2 show the locations of the polling stations in the European election compared to the locations of maximum and minimum voter density and Figures 3 & 4 show the same for the local election.

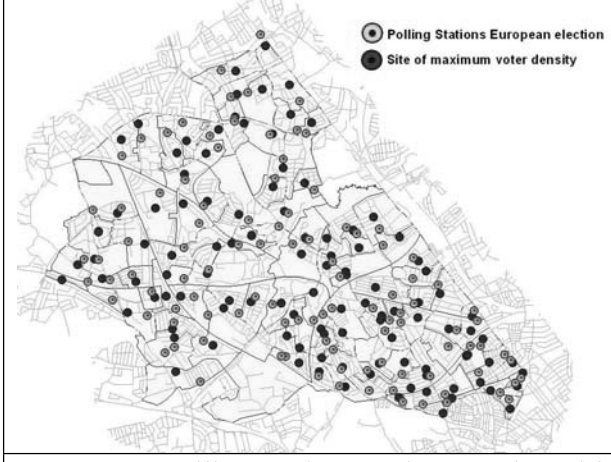



**Figure 1.** Polling stations and postcodes with the maximum voter density in the European

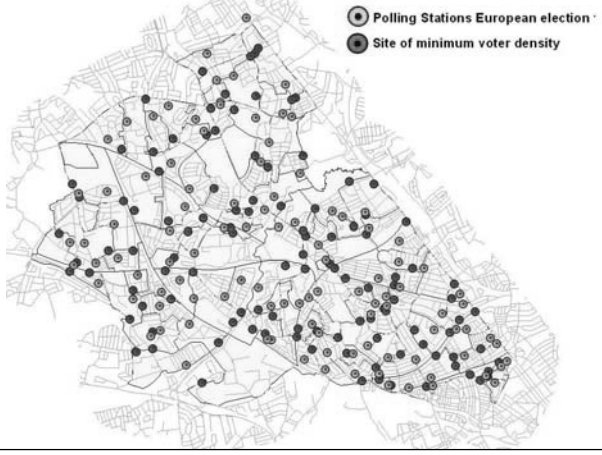
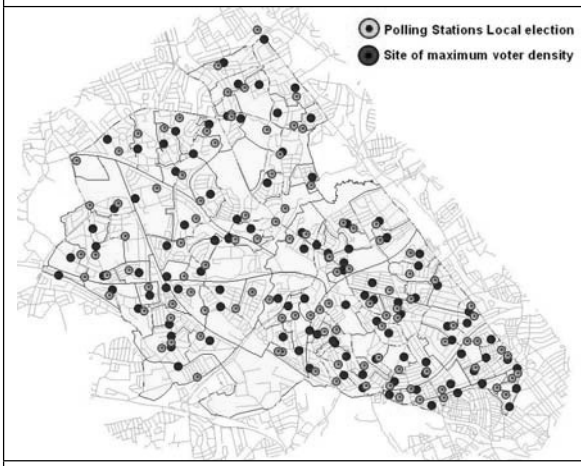
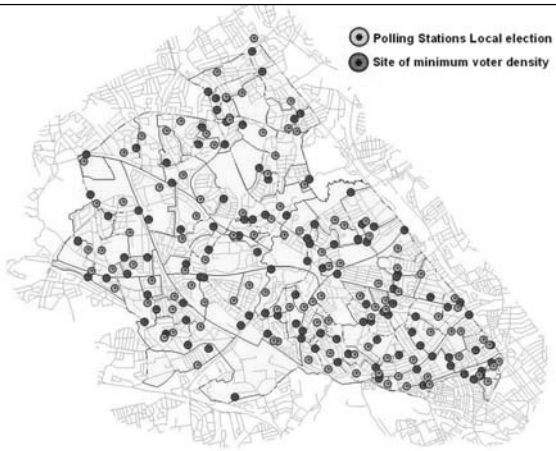



**Figure 2.** Polling stations and postcodes with the minimum voter density in the European

| election | election |
|---|---|
|   |   |
| **Figure 3.** Polling stations and postcodes with the maximum voter density in the local election | **Figure 4.** Polling stations and postcodes with the minimum voter density in the local election |

## 4. Statistical analysis

Three binomial multi-level models were estimated for each of the three elections across the three spatial scales of analysis (polling district, ward and constituency). Proportion voter turnout per polling district was the dependent variable and a variety of social and spatial measures generated from census and GIS data that are known to affect turnout (such as the voter density variables) were included as independent variables (Orford et al., 2009). These are shown in Table 1. Published literature suggested the important socio-economic variables that affect turnout and these were measured using output area level census data re-apportioned to polling districts. Factor Analysis was used to collapse the census data into a smaller number of uncorrelated variables that were interpreted as representing relative deprivation, the proportion of the student population, the proportion of the retired population and the proportion of white people in professional occupations in a polling district.

The results showed that voter density did not have a statistically significant impact on turnout to parliamentary elections but did have an important and significant effect in local and European elections. The deprivation variable had significant negative effect in all three elections, with the student population variable only being significant in the European election and the retired population variable having a significant effect in the local and Parliamentary elections. The terrain variable had a significant negative effect in the European election and the marginality variable had a significant positive effect in the local election. The variables measuring polling station building type and the status of the road on which it was located were not significant in any of the elections.

**Table 1.** Independent variables in the model

| Variables | Type |
|---|---|
| Relative deprivation of polling district | Continuous |
| Student population | Continuous |
| Retired population | Continuous |
| White professional population | Continuous |
| Difference in topographical height between postcode and polling station (Terrain) | Continuous |
| Voter density of postcode | Continuous |

| | |
|---|---|
| Marginality of ward (majority of ten percentage points or less - only used in local election model) | Dummy |
| Polling station building type (school, hall, library, community centre, portable cabin, other) | Dummy |
| Road status of polling station location (A-Road, B-Road, minor road) | Dummy |

The multi-level models for local and European elections were then used to predict the changes in the probability of turnout if the polling stations were re-sited at the postcodes of the maximum, minimum and average voter densities. The differences at polling district level are shown graphically in figure 5 for the European election and figure 6 for the local election. For 10% and 25% of polling districts in European and local elections respectively there is no substantial difference in the rate of predicted turnout from moving the polling station to the location of maximum voter density. Here, the polling stations are already in or close to the optimal location and so will not benefit from moving. In comparison, around 10% of polling stations in European elections and 2% in the local election had an increase in predicted probability of turnout of 3% or more. In terms of re-siting the polling station to the location of the lowest voter density, 50% of polling stations have at least a 2% fall in predicted probability of turnout in the European and local elections. For 10% of polling stations in the European election (and 5% in the local election) this drop in predicted probability of turnout is more than 3.5% and it is over 4.5% for one polling district.

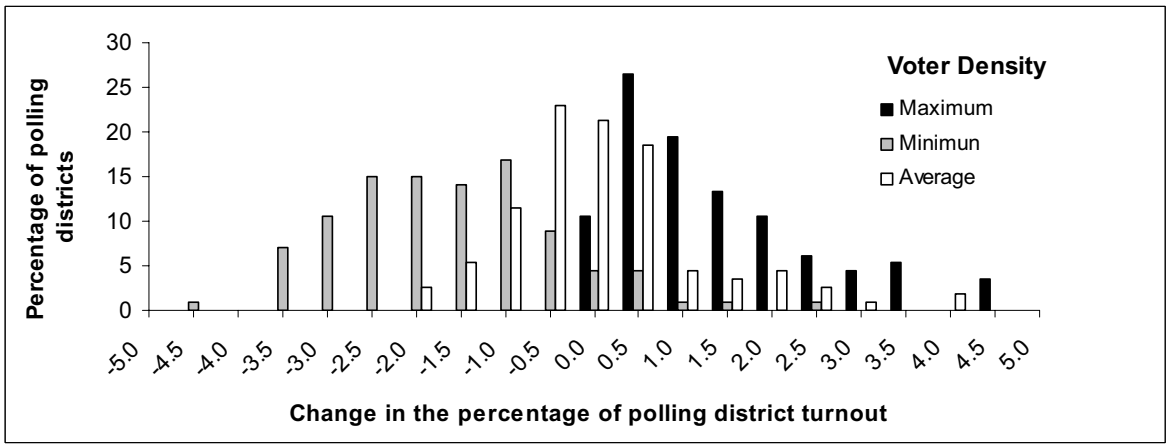



**Figure 5:** Percentage differences in the predicted probability of turnout at polling district level when re-siting polling stations in the European election

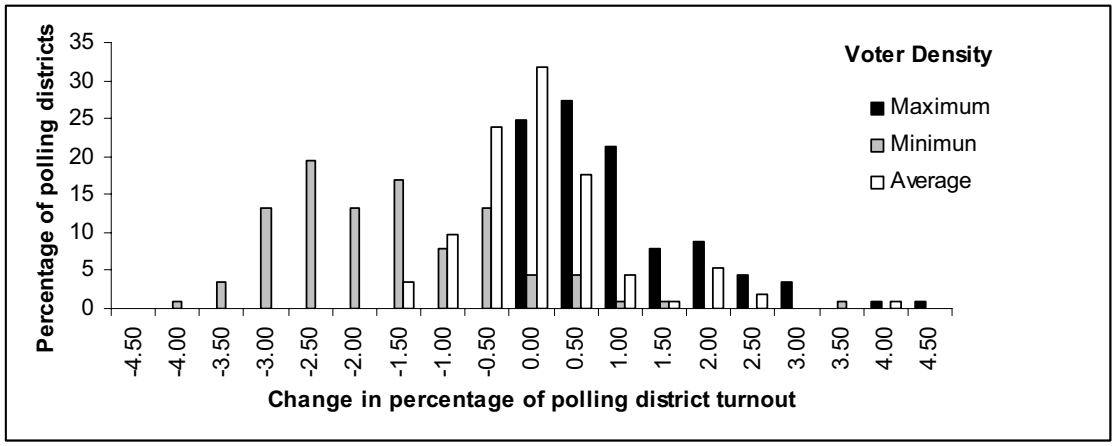



**Figure 6:** Percentage differences in the predicted probability of turnout at polling district

level when re-siting polling stations in the local election

## 5. Conclusion

For many polling districts the increases in the predicted probability of turnout by re-siting a polling station is rather small. However, this is to be expected given the small area of most polling districts and the fact that many polling stations are already located close to the point of maximum voter density. But by drilling down into ward and polling district level analysis, it becomes clearer that the effect of re-siting becomes significant for a small number of polling stations, with differences of 3%-4% in the predicted probability of turnout between the actual and optimal polling station locations. These differences can increase to 4%-5% when comparing the most and least optimal locations for a polling station within a polling district. Therefore, for a small number polling districts, re-siting the polling station to a more optimal location with respect to voter density may be appropriate in order to increase turnout. Also, in city like London, people's preferences for polling station locations may also be determined by proximity to public transport nodes. A factor to consider here is whether a suitable premise exists at the optimal location, given that this constraint was relaxed in our analysis. A local authority can always use a portable polling station but there are several considerations that will need to be planned for, including making road closure orders if necessary, providing toilet facilities and electricity. This can increase the cost and bureaucracy of an election and therefore should be used only if predicted turnout will be significantly increased.

## References

Electoral Commission (2007a) *Electronic Voting: May 2007 electoral pilot schemes*, London: The Electoral Commission

Electoral Commission (2007b) *Handbook for polling station staff – supporting local government elections in England and Wales*, London: The Electoral Commission.

Electoral Commission (2007c) *Managing a local government election in England and Wales: guidance for Returning Officers, a good practice guidance manual*, London: The Electoral Commission.

Schuknecht J E 2003 Political participation and the accessibility of the ballot box *Political Geography* **22** pp471–88

Haspel M and Knotts G 2005 Location, location, location: precinct placement and the costs of voting *Journal of Politics* **67** pp560–73

Niemi R G, Grofman B, Carlucci C and Hofeller T 1990, Measuring compactness and the role of a compactness standard in a test for partisan and racial Gerrymandering, *Journal of Politics* 52 pp1155–81

Orford, S., Rallings, C., Thrasher, M., and Borisyuk, G., (2009) Electoral salience and the costs of voting at national, sub-national and supra-national elections in the UK: a case study of Brent, UK, *Transactions Institute of British Geographers*, 34 pp195–214

Orford, S., Rallings, C., Thrasher, M., and Borisyuk, G., (2008) Investigating differences in electoral turnout: the influence of ward-level context on participation in local and parliamentary elections in Britain, *Environment and Planning A*, 40, pp1250 – 1268

Orford, S., and Schuman, A., (2002) Micro-geography and socialization: new ways of investigating the turnout problem" in Bennie L., Rallings C., Tonge J., and Webb, P. (eds) *British elections and parties review 12: the 2001 general election* Frank Cass, London 190-205

## Biography

*Scott Orford is a senior lecturer in Spatial Analysis and GIS in the School of City and Regional Planning, Cardiff University and the director of the WISERD Data Team. His research interests include the analysis and modelling of socio-economic processes, innovation in data integration and*

*data construction and the mapping and spatial analysis of qualitative data*

# A containment-first search algorithm for higher-order analysis of urban topology

J.-P. de Almeida[a, b, 1], J. G. Morley[c], I. J. Dowman[d]

[a] Geomatic Engineering Unit, Dept of Mathematics, Faculty of Science and Technology, University of Coimbra (FCTUC), Largo D. Dinis, Apartado 3008, 3001-454 COIMBRA, Portugal
Tel. +351 239 791 150 Fax +351 239 832 568
Email: zepaulo@mat.uc.pt | http://www.mat.uc.pt

[b] Institute for Systems Engineering and Computers at Coimbra (INESCC),
Rua Antero de Quental, 199, 3000-141 COIMBRA, Portugal
Tel. +351 239 851 040/9 Fax +351 239 824 692
http://kreation.dec.uc.pt/projects

[c] Centre for Geospatial Science, School of Geography, University of Nottingham (UN),
Sir Clive Granger Building, University Park, NOTTINGHAM NG7 2RD, UK
Tel. +44 (0)11 5846 8411 Fax +44 (0)11 5951 5249
jeremy.morley@nottingham.ac.uk
http://www.nottingham.ac.uk/geography/CGS/

[d] Dept of Civil, Environmental & Geomatic Engineering, University College London (UCL),
Gower Street, LONDON WC1E 6BT, UK
Tel. +44 (0)20 7679 2740 Fax +44 (0)20 7380 0453
idowman@cege.ucl.ac.uk
http://www.cege.ucl.ac.uk

**Keywords**: topology, Graph Theory, urban; scene analysis, GIS.

## 1. Introduction

### 1.1 Motivation

Research has revealed the importance of the concepts from the mathematical areas of both topology and graph theory for interpreting the spatial arrangement of spatial entities. Graph theory in particular has been used in different applications of a wide range of fields for that purpose, however not many graph-theoretic approaches to analyse entities within the urban environment are available in the literature. Some examples should be mentioned though such as, Bafna (2003), Barr and Barnsley (2004), Bunn *et al*. (2000), Krüger (1999), Nardinochi *et al*. (2003), and Steel *et al*. (2003).

Very little work has been devoted in particular to the interpretation of initially unstructured geospatial datasets. In most of the applications developed up-to-date for the interpretation and analysis of spatial phenomena within the urban context, the starting point is to some extent a meaningful dataset in terms of the urban scene. Starting at a level further back, before meaningful data are obtained, the interpretation and analysis of spatial phenomena are more challenging tasks and require further investigation.

The aim of retrieving structured information from initial unstructured spatial data, translated into more meaningful homogeneous regions, can be achieved by identifying meaningful structures within the initial random collection of objects and by understanding their spatial arrangement (Anders *et al.*, 1999). It is believed that the task of understanding topological relationships between objects can be accomplished by both applying graph theory and carrying out graph analysis (de Almeida *et al.*, 2007).

### 1.2 Background

Starting from initially unstructured geospatial datasets of urban areas (thus, no prior knowledge of the spatial entities is assumed), de Almeida *et al.* (2005, 2007) showed how a graph-theoretic approach could be applied towards the analysis of the urban scene spatial topology.

Urban LiDAR data was used as an example scenario. Topology was initially brought in to the original data by generating a triangulated irregular network (TIN - the maximal planar description of the given point set's internal structure, Kirkpatrick and Radke, 1985). A binary classification of the TIN facets based upon their gradient – whose thresholding depends on the resolution of the initial data – was employed ("flat" and "steep" facets). Eventually, the TIN facets were "aggregated" according to the classification above that led to a map of polygonal gradient regions ("flat" and "steep" polygons). The authors pointed out how the steep polygonal regions in particular were expected to enclose urban features. These steps constitute the preliminary preparation process of raw data.

A network of connectivity throughout the map of flat & steep polygonal regions was then built up by applying graph theory, which resulted in a graph of adjacencies: each region in the graph is represented by a node; graph edges link up nodes corresponding to adjacent polygons. The adjacency graph was processed either through the depth-first (DFS) or the breadth-first search (BFS) algorithms. Given the different ways each algorithm operates in traversing a graph, it was noted how BFS results are more meaningful in terms of the urban scene: the BFS tree branches are connected components of the original graph, and represent the shortest path between the root and their leaf (Sedgewick, 2002); it seems that they can be related to potential urban features. Thus, the implementation of the graph analysis procedure was based upon BFS. It traverses the graph looking for sequential relationships of containment amongst the sequences of adjacency: containment-first search (CFS). In fact, where containment occurs within the Useful External Border (UEB) – basically, the outer flat enclosing polygon corresponding to the ground – there is a high likelihood of an urban feature being present (de Almeida *et al*., 2005, 2007).

This paper describes in detail CFS algorithm which was developed for the purposes above. The diagram depicted in Figure 1 above illustrates where the algorithm sits within the whole methodology proposed for the analysis of urban spatial topology.

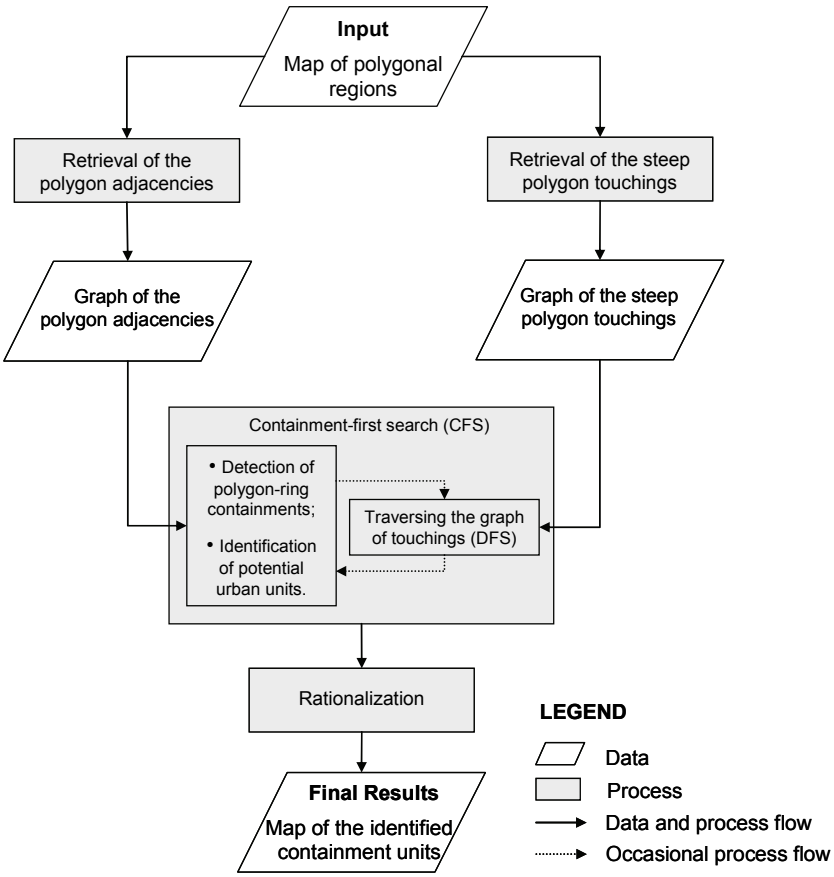



**Figure 1**. An overview of the methodology proposed for the analysis of urban spatial topology (de Almeida, 2007).

## 2. A containment-first search algorithm

### 2.1 Preliminaries

For the purposes of this work, when two polygons share at least one arc, the spatial relation is called *adjacency*; if the two polygons happen to meet at a node, the spatial relation is distinguished from the previous one and is called *touching* (de Almeida *et al.*, 2007; de Almeida, 2007).

As de Almeida *et al.* (2007) noted, CFS could not be developed simply based on BFS but had to be extended in order to be able to detect the spatial relation of containment in a broader sense. The spatial relation of *touching* between steep polygons should be taken into consideration. This improvement enabled the derivation of particular cases of containment not explicit in the graph of adjacencies, e.g. when a ring of steep polygons meeting at nodes contains a single flat polygon - the so-called polygon-ring containments.

### 2.2 The analytical analysis method

The graph-theoretic approach was implemented based on the investigation of the topological relationships between objects in the context of the whole spatial scene. This was coded in C foreseeing the advantages and potentialities of pointer structures in C for graph analysis (Kelley and Pohl, 1990).

The analysis method can be interpreted as follows. Considering the UEB (recall that this is a flat polygon) as the starting point of the search process, the original graph of adjacencies is

traversed. When visiting the adjacent steep vertices of the root, the CFS algorithm takes the first vertex appearing in the root's linked list and, starting from this one, traverses the graph of steep-polygon touchings. Because the graph of touchings is a disconnected graph, the traversal process covers only the subgraph that the given steep vertex belongs to. While traversing this particular subgraph, the CFS algorithm tags all the steep vertices visited as belonging to the same connected unit. This process continues until the first level of adjacency of the graph of adjacencies is exhausted. When the CFS comes across a root's adjacent vertex already tagged as belonging to a particular containment unit, this is skipped and the corresponding polygon remains intact, belonging to the containment unit already identified.

To illustrate the concept implemented, let us take a simpler scene pictured in Figure 2. Let us suppose that steep polygons *3,...,11* (in dark green) are constituent parts of the rings of steep polygons enclosing flat polygons *12* and *13* (*vd.* Figure 2a); in other words, there is a sub-graph of the graph of steep-polygon touchings that consists of vertices *3* to *11*.

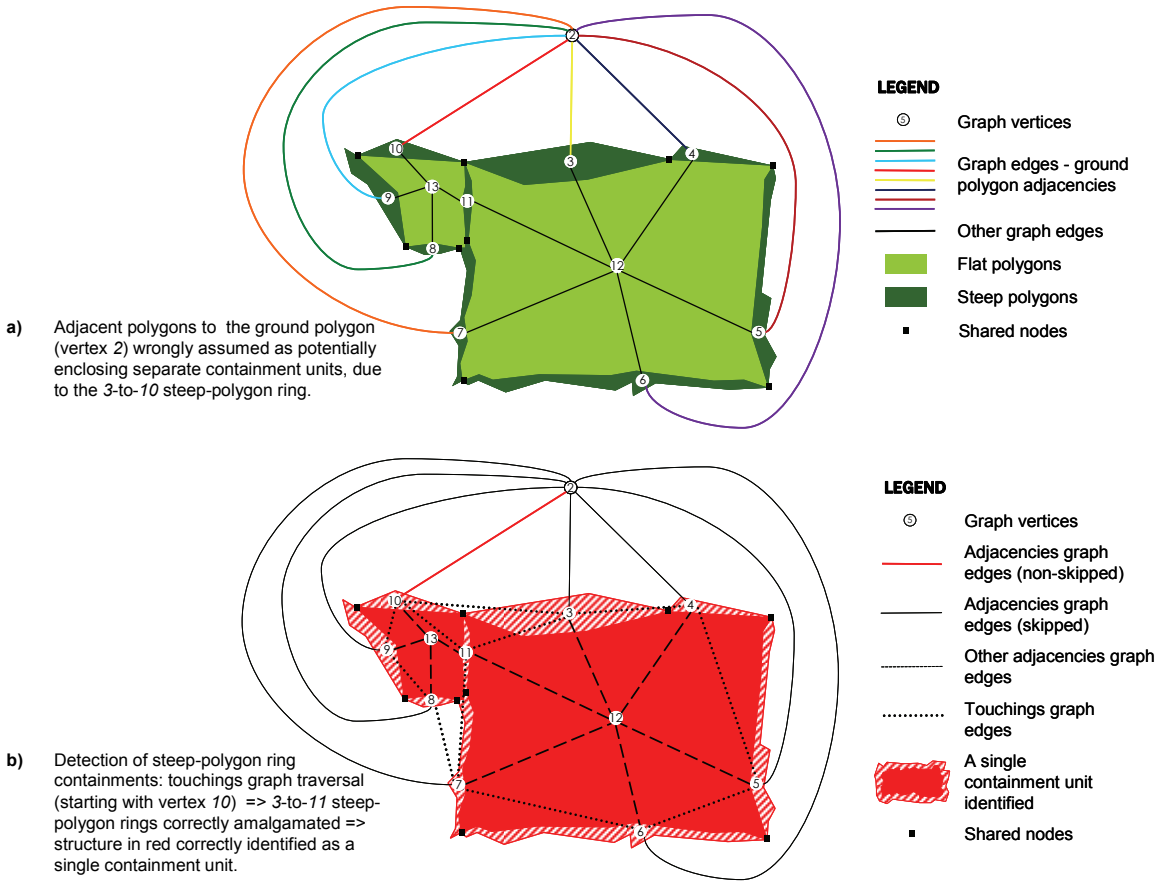



**Figure 2**. The containment-first search process: **a)** before polygon-ring containments are detected; **b)** after polygon-ring containments are detected

When vertex *2* is visited in the adjacencies graph, the algorithm takes the vertex at the top of *2*'s adjacency list, vertex *10*, and the graph of steep-polygon touchings is traversed starting from *10*; all the steep vertices belonging to the same sub-graph as that of *10* are tagged accordingly, indicating a potential containment unit. When vertex *10* is exhausted in the graph traversal, CFS moves on to visit vertex *9*; this is now skipped since it was previously tagged as belonging to an containment unit already identified. And so on so forth until vertex *3* is visited, and the containment unit is complete. Visually, the translation of the facts above is accomplished by assigning the same colour to all steep (hashed pattern) and flat (solid colour) polygons within the same containment unit (*vd.* Figure 2b).

## 3. Proof of concept

### 3.1 Generation of synthetic data

Before tests with real initial unstructured urban data of are undertaken, this section describes an experiment carried out with synthetic idealised spatial data relating to urban objects. A map of binary classified gradient regions was created simulating a map of higher-level urban scene objects. This was derived from building polygons from OS Master Map data[2].

As Figure 3 shows, steep polygons shape both buildings standing on their own and higher-level structures; the enclosed flat polygons simulate building roofs. For topological reasons, an outer polygon - distinct from the Universe Polygon - had to be considered so as to simulate the ground polygon.

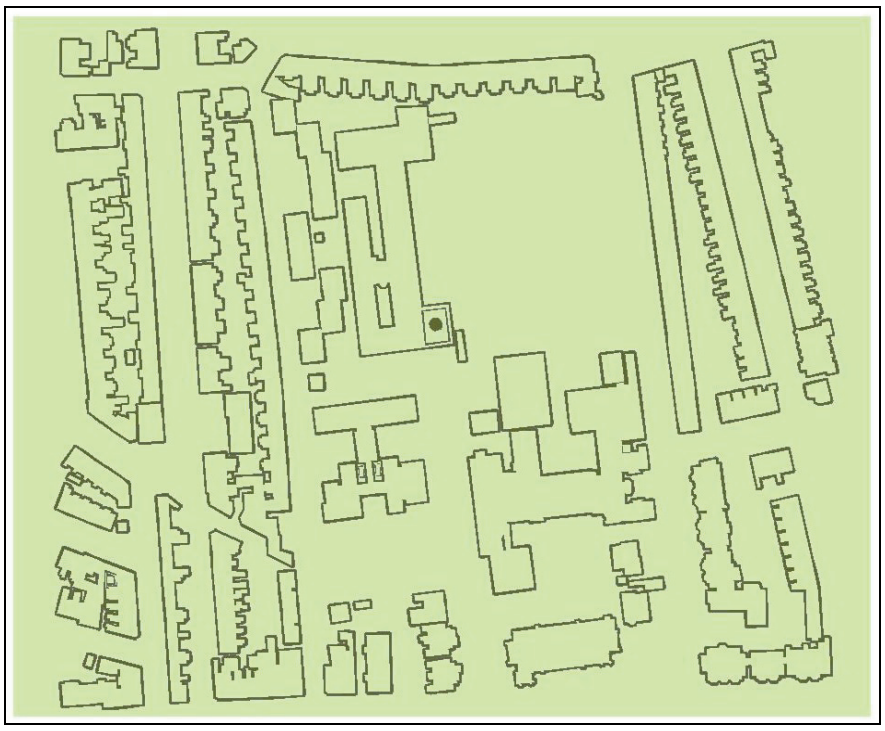


**Figure 3**. Simulated map of gradient regions: binary classification of the polygons generated into "steep" polygons (dark colour) and "flat" polygons (light colour).

Polygon and associate arc attributes were accessed in order to retrieve gradient-region adjacencies (ESRI, 1995; Rigaux *et al*., 2002; ESRI, 2005). The graph of adjacencies was generated. The main characteristic of this graph when dealing with clean data is the fact that all relationships of adjacency are also of containment. This is far more complex when dealing with real world data: it is not guaranteed beforehand that a spatial relation of adjacency is also of containment (de Almeida *et al.*, 2007).

As typically happens with real world data, there are no steep polygons meeting at nodes in this case (*i.e.* steep-polygon rings, enclosing single containment units, are not split into different entities), and hence the touchings graph is a null graph.

[2] Made available by the Department of Geomatic Engineering of the University College London for academic purposes. Ordnance Survey ©Crown Copyright, all rights reserved.

### 3.2 Spatial topology analysis

Figure 4 depicts the results of the spatial topology analysis for the simulated map of gradient regions. Polygon *2* (corresponding to the ground polygon, mapped in white), with 45 adjacent regions, was chosen as the UEB.

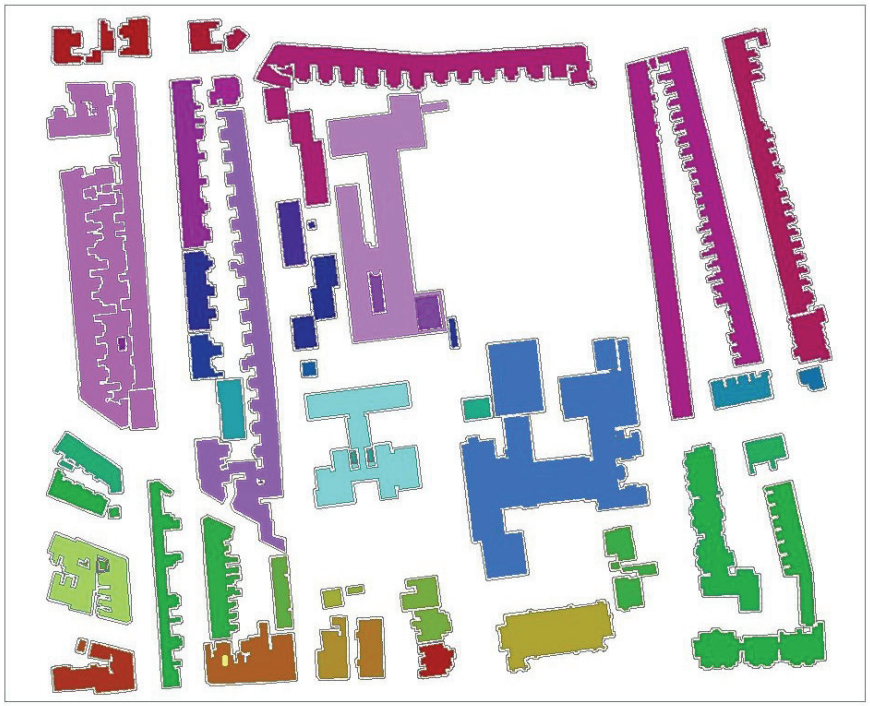


**Figure 4**. Spatial topology analysis and the different containment units identified.

It can be seen that the algorithm indeed detected individually all the urban features simulated as separate containment units (different colours represent each one). In fact, sequences of adjacencies/containments were correctly detected as so by the algorithm (*vd*. Figure 4): solid colours correspond to flat polygons; coloured hashed patterns correspond to steep polygons. Moreover, individual simulated spatial features, closely standing next to one another but not actually juxtaposed, were detected separately. This confirms that in theory the algorithm is even capable of detecting single buildings standing on their own.

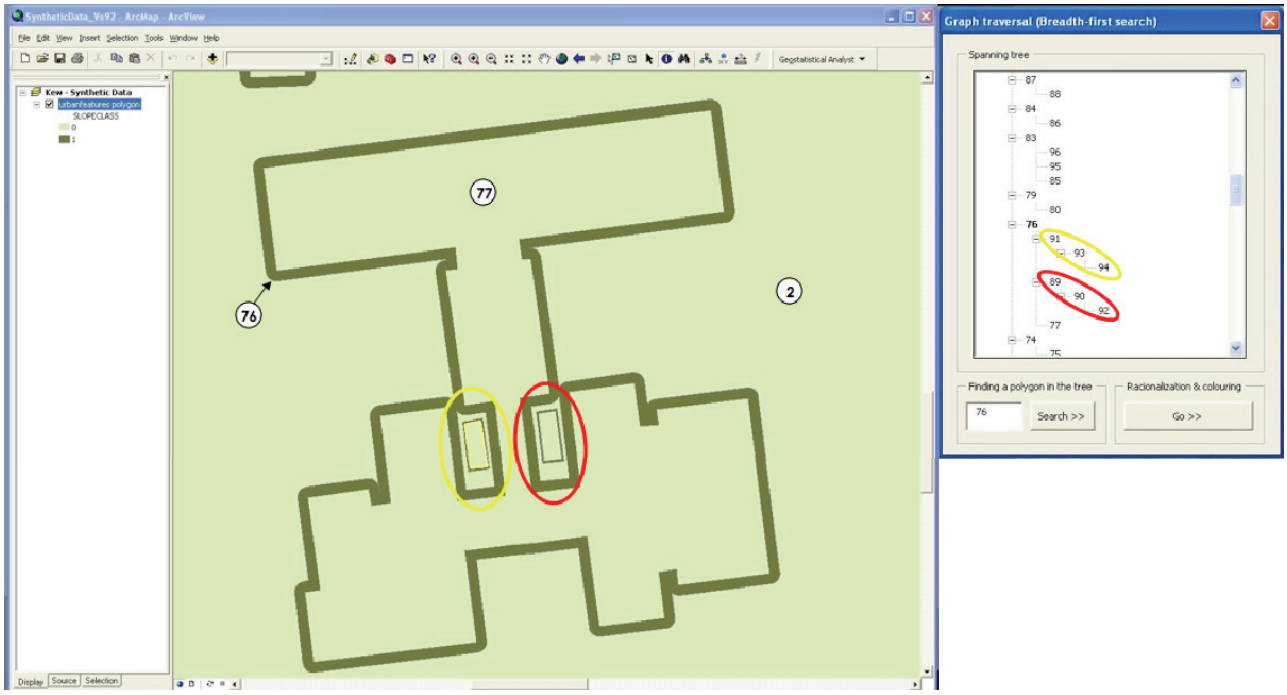



**Figure 5**. Branches of the breadth-first search tree directly relate to spatial features. (Detail of Figure 3; the numbers on the map are polygon labels in insert)

## 4. Summary and conclusions

Further to our prior work, this paper showed how the spatial relation of touching between steep polygons was taken into consideration in order to extend the CFS procedure to be able to detect polygon-ring containments. A flowchart of the algorithms implemented was provided, and an illustration of how CFS procedure works was also given.

Proof of concept was carried out. For the purpose, synthetic data was generated and a map of gradient regions, simulating urban scene objects, was created. The analysis of the spatial topology was undertaken, and conclusions were drawn in terms of the assertions made when designing the algorithms. The results obtained demonstrated that, in the absence of noise and error, the algorithms do indeed make the urban spatial topology more explicit. In particular, the results support the assumption that each BFS tree's branch does relate to a single containment unit within the initial map of gradient regions (*e.g.* tree branch starting at vertex *76* in Figure 5). Moreover, sequences of containment relationships do relate to higher-level urban scene objects.

The concepts drawn in the research described and the algorithm implemented should serve as the basis for automatic analysis of spatial datasets, such as: analysis of image data; analysis of settlement structures; automation of land use mapping for urban areas. Furthermore, the results obtained particularly with LiDAR data reveal that the methodology proposed has also promised as a tool that could be extended to be applied in the automatic classification of raw LiDAR data.

## References

de ALMEIDA J.-P., MORLEY J. G., DOWMAN I. J. (2005). A graph-based technique for higher order topological data structure visualisation. In *Proceedings of GISRUK 2005*: 306-312.

de ALMEIDA J.-P., MORLEY J. G., DOWMAN I. J. (2007). Graph theory in higher order topological analysis of urban scenes. In *Computers, Environment and Urban Systems* **31**(4): 426-440.

de ALMEIDA J.-P. (2007). *A Graph-based Technique for Analysis and Visualisation of Higher Order Urban Topology*. PhD thesis (unpublished). Department of Geomatic Engineering, University College London, London (UK).

ANDERS, K.-H.; SESTER, M.; FRITSCH, D. (1999). Analysis of settlement structures by graph-based clustering. *Semantische Modellierung, SMATI* **99**: 41-49, Munich, (Germany).

BAFNA, S. (2003). Space Syntax – A brief introduction to its logic and analytical techniques. *Environment and behaviour* **35**(1): 17-29. Sage Publications.

BARR, S.L., BARNSLEY, M.J. (2004). Characterising the structural form of an urban system using built-form connectivity model concepts. *Spatial Modelling of the Terrestrial Environment* (KELLY, R.E.J., DRAKE, N.A., BARR, S.L., Eds.): 201-226. John Wiley, Chichester (England, UK).

BUNN, A.G., URBAN, D.L., KEITT, T.H. (2000). Landscape connectivity: A conservation application of graph theory. *Journal of Environmental Management* **59**: 265-275. Academic Press (USA)

ESRI (1995). *Understanding GIS, The ARC/INFO Method*, Lesson2. ESRI, Inc. Cambridge, England (UK).

ESRI (2005). ArcGIS Topology [online]. Available from http://www.esriuk.com/products/support/ArcGIS%20UK%20EXT/Topology.asp [Accessed 10 June 2005].

KELLEY A., POHL I. (1990). *A Book on C - Programming in C* (2nd Edn.). Benjamin Cummings, Redwood City, CA (USA).

KIRKPATRICK, D.G., RADKE, J.D. (1985). *A framework for computational morphology. Computational Geometry* (TOUSSAINT, G.T., Ed.): 217-248. North-Holland, Amsterdam (The Netherlands).

KRÜGER, MÁRIO (1999). On node and axial maps: distance measures and related topics. *European Conference on the Representation and Management of Urban Change*: 28-29 September 1999. Cambridge (England, UK).

NARDINOCHI, C., FORLANI, G., ZINGARETTI, P. (2003). Classification and filtering of laser data. *ISPRS Workshop on 3D reconstruction from airborne laser scanner and InSAR data*. Dresden (Germany).

RIGAUX J., SCHOLL M., VOISARD A. (2002). *Spatial Databases with Applications to GIS*, **8**. Morgan Kaufmann, San Francisco, CA (USA).

SEDGEWICK. R. (2002). *Algorithms in C, Part 5* (3rd Edn.), **17**, **18**. Addison-Wesley, Reading, MA (USA).

STEEL, A.M., BARNSLEY, M.J., BARR, S.L. (2003). Inferring urban land use through analysis of the spatial composition of buildings identified in LiDAR and multispectral image data. *Remotely-Sensed Cities* (MESEV, T.V., Ed.). Taylor & Francis, London (England, UK).

## Acknowledgements

*Faculdade de Ciências e Tecnologia da Universidade de Coimbra* (Faculty of Science & Technology of the University of Coimbra, Portugal), and *Fundação para a Ciência e Tecnologia* (The Science & Technology Foundation – Portuguese Ministry of Science,

Technology and Higher Education - Scholarship SFRH/BD/9909/2002), which supported financially this research.

An acknowledgement is also due to Dr. João Coutinho-Rodrigues (Department of Civil Engineering, Faculty of Science and Technology, University of Coimbra, Portugal), J-P de Almeida's PhD co-supervisor at the University of Coimbra.

## Biography

***José-Paulo de Almeida*** *(Grad Geomatic Engineering UC; MSc Civil Engineering - Specialisation Urban Engineering UC; PhD Geomatic Engineering UCL) has been working at UC where he is currently Lecturer of Geomatic Engineering; he is also a junior researcher at INESCC. He's been working on: the interpretation of unstructured geospatial data in GIS environment using Graph Theory; spatial decision support systems; currently he's also interested on the semantic enrichment of 3D data towards the development of 3D city models.*

***Jeremy Morley*** *(MA Natural Sciences U. Cambridge; MSc Remote Sensing UCL & Imperial C) is Deputy Director of the Centre for Geospatial Science (CGS) at the University of Nottingham. He was programme director of UCL's MSc in GIS (1998-2004) and of its BEng/MEng in Geoinformatics (2005-9). Over the last 15 years his research has focussed on the mapping of Mars in support of geological analysis; terrain mapping from LiDAR and InSAR; GIS interoperability and mashup WebGIS systems. He's been UCL's technical representative to the Open Geospatial Consortium since 2004.*

***Ian Dowman*** *(BSc Geography UCL; PhD Photogrammetry UCL; FRICS - Fellow of the Royal institution of Chartered Surveyor) has spent most of his career at UCL, where he's currently Professor of Photogrammetry and Remote Sensing. He's been involved in ISPRS for many years as a working group chair, commission president, and as Secretary General of ISPRS. He's also been chair of the sub group on Terrain Mapping of the CEOS Working Group on Calibration and Validation. He worked with aerial and close range photography, on the application of satellite data to mapping, and is currently involved in automatic feature extraction and geometric fusion of different types of data.*

# Street-level Point Interpolation

Narushige Shiode[1], Shino Shiode[2]

[1]School of City and Regional Planning – Cardiff University
Glamorgan Building, King George VII Avenue, Cardiff CF10 3WA, UK
Email: ShiodeN@cardiff.ac.uk

[2]Department of Geography, Environment and Development Studies – Birkbeck, University of London
Malet Street, London WC1E 7HX, UK
Email: s.shiode@bbk.ac.uk

KEYWORDS: cross validation, Inverse Distance-Weight, Network, Shortest-Path, Spatial Interpolation

## 1. Introduction

Spatial interpolation is a method commonly used for estimating an unknown spatial value using known values observed at a set of sample locations. Cresssie (2003) states that the need to obtain a good estimate of such a value can be found in nearly all scientific disciplines. Applications are indeed found in a number of disciplines including climatology, geostatistics, geomorphology and environmental study. A variety of spatial interpolation methods exist, each providing a good prediction under different estimation criteria. In recent years, these methods have become an integral part of GIS.

Existing methods assume that the all locations exist in the Euclidean space, i.e. the distance between the samples and the targets is measured in a straight line. Although this assumption holds in many cases, some phenomena are observed or measured on a network, which should be analysed using the network space (e.g. the elevations of a terrain surface, the appraisal value of estate properties and the gas emissions level along highways). This study proposes a network-based spatial interpolation method for estimating unknown values at locations along a network. It extends an existing discrete, local interpolation method and adapting it to the network space.

## 2. Network-based Inverse Distance-Weighted Interpolation

We introduce a method that predicts an unknown spatial value on a network using observed values at nearby sample locations and weight it with respect to the shortest-path distance from them. We will hereafter call it the network inverse distance-weighted method (NT-IDW) and distinguish it from the conventional IDW, which we will refer to as the planar inverse distance-weighted method (PL-IDW). The NT-IDW is carried out in three steps: (1) find, with a shortest-path search, a fixed number of points closest to the target location and identify them as the nearby sample locations; (2) calculate the weight as an inverse power function of the shortest-path distance between each sample location and the target location; and (3) predict the unknown value as the weighted mean of the observed values.

Let $p_0$ be the target location on the network, and $z_0$ be the unknown value of $p_0$. Let $p_1, p_2, \ldots, p_s$ be $s$ number of nearby sample locations with observed values of $z_1, z_2, \ldots, z_s$, respectively. Then $z_0$ is predicted as the weighted mean of the observed values at the nearby sample locations:

$$\hat{z}_0 = \sum_{i=1}^{s} w_i z_i \quad , \quad \sum_{i=1}^{s} w_i = 1 \tag{1}$$

where $W = w_i$ ($i$=1,2,...,$s$) is the weight assigned to $p_1, p_2, \ldots, p_s$ relative to the shortest-path distance from $p_0$. Let $d_{NT}$ ($p_0$, $p_i$) ($i$=1,2,...,$s$) denote the shortest-path distance from a nearby sample location $p_i$ ($i$=1,2,...,$s$) to the target location $p_0$ along the network. Equation 1 can be rewritten as

$$\sum_{i=1}^{s} w_i z_i = \frac{\sum_{i=1}^{s} f(d_{\mathrm{NT}}(p_0, p_i))z_i}{\sum_{i=1}^{s} f(d_{\mathrm{NT}}(p_0, p_i))} \tag{2}$$

The influence of the observed value at the nearby sample locations on the unknown value at the target location is assumed to weaken as the distance between them increases. In order to account for the effect of distance decay, the inverse power function of $d_{\mathrm{NT}}$ is introduced as follows

$$\hat{z}_0 = \frac{\sum_{i=1}^{s} d_{\mathrm{NT}}(p_0, p_i)^{-\lambda} z_i}{\sum_{i=1}^{s} d_{\mathrm{NT}}(p_0, p_i)^{-\lambda}} \tag{3}$$

where $\lambda$ is the power coefficient.

## 3. Inverse Distance-Weighted Interpolation in Planar and Network Spaces

The validity of NT-IDW can be confirmed by using a cross-validation technique on both PL-IDW and NT-IDW. Suppose that there are $n$ number of sample locations $p_1, p_2,..., p_n$ in the study area, and that the observed values at each sample location are denoted by $z_1, z_2,..., z_n$. Let the predicted attribute values for NT-IDW and PL-IDW be $\hat{z}_i^{\mathrm{NT}}$, $\hat{z}_i^{\mathrm{PL}}$ for target locations $p_i$ ($i$ = 1, 2,..., $n$). An index for comparing the prediction error of $\hat{z}_i^{\mathrm{NT}}$, $\hat{z}_i^{\mathrm{PL}}$ from the observed value $z_i$ can be defined as

$$r_i^{\mathrm{NT}} = \hat{z}_i^{\mathrm{NT}} - z_i \ , \ r_i^{\mathrm{PL}} = \hat{z}_i^{\mathrm{PL}} - z_i \ , (i = 1, 2,..., n). \tag{4}$$

Then, Mean Squared Errors (MSE) (Isaaks and Srivastava 1989) for PL-IDW and NT-IDW are

$$M_{\mathrm{NT}} = \frac{1}{n}\sum_{i=1}^{n}(r_i^{\mathrm{NT}})^2, M_{\mathrm{PL}} = \frac{1}{n}\sum_{i=1}^{n}(r_i^{\mathrm{PL}})^2 \ . \tag{5}$$

In addition to MSE, which is an aggregated index, we introduce index $D$ to show the difference in the degree of disparity of the prediction error between NT-IDW and PL-IDW for each location $p_i$ as

$$D_i = \left|r_i^{\mathrm{NT}}\right| - \left|r_i^{\mathrm{PL}}\right|, (i = 1, 2,..., n). \tag{6}$$

## 4. Comparative Study of the Network and the Planar Interpolations

### 4.1 Interpolated Results from Data 1

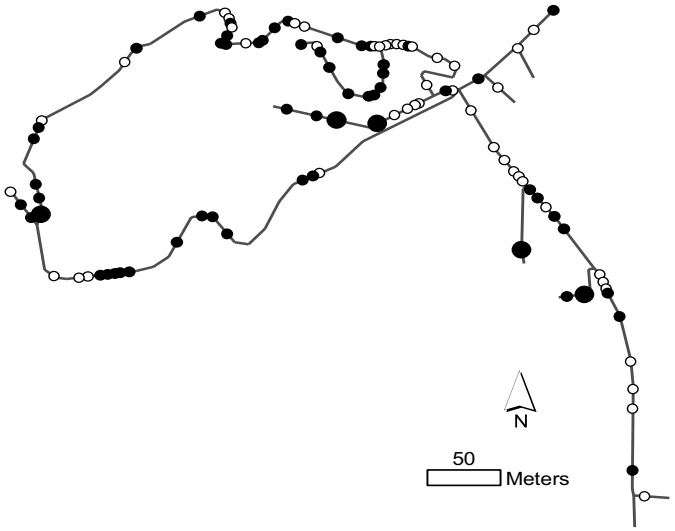


**Figure 1.** A street network along a ridge line containing 104 sample locations with an illustration of the relative prediction accuracy of NT-IDW at each location. Black circles show target locations at which NT-IDW outperformed PL-IDW. Small circles show those with average $D$ value ($-2<D<0$) and large circles with smaller $D$ value ($D<-2$) where NT-IDW achieved much higher prediction accuracy.

In order to perform the cross-validation, we adopt two sets of elevation data measured on a network. Data 1 is a street network with 104 sample locations that follows a simple but winding ridge-line (elevation level: 28~84 m) (Figure 1). Table 1 shows the result of the cross-validation calculated at all sample locations in Data 1 with the parameters taking a range of values at $s$ = 2, 3,…, 8, $\lambda$ = 0.5, 1,…, 3. The upper and the lower section of each cell show MSE of NT-IDW and PL-IDW, respectively (the smaller value is underscored to indicate the one with higher accuracy). They confirm that NT-IDW generally maintains a higher degree of prediction accuracy for all combinations of $s$ and $\lambda$. In addition, $D$ index measures the degree of disparity between the prediction errors from the two methods at each individual location. In Figure 1, the sample locations at which NT-IDW outperformed PL-IDW are shown in black circles.

**Table 1.** MSEs of the planar and the network interpolations of Data 1.

| $\lambda$ \ $s$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 0.5 | **5.924** | **6.236** | **7.323** | **8.977** | **11.012** | **12.104** | **13.916** |
| | 7.874 | 9.381 | 10.767 | 12.919 | 14.701 | 16.512 | 17.252 |
| 1 | **5.479** | **5.484** | **6.265** | **7.178** | **8.429** | **9.019** | **10.152** |
| | 7.449 | 8.481 | 9.173 | 10.512 | 11.562 | 12.560 | 13.089 |
| 1.5 | **5.162** | **4.991** | **5.529** | **5.968** | **6.698** | **6.956** | **7.595** |
| | 7.174 | 7.843 | 8.070 | 8.840 | 9.393 | 9.866 | 10.161 |
| 2 | **4.951** | **4.689** | **5.062** | **5.232** | **5.651** | **5.733** | **6.082** |
| | 7.019 | 7.413 | 7.372 | 7.790 | 8.049 | 8.252 | 8.403 |
| 2.5 | **4.818** | **4.514** | **4.781** | **4.813** | **5.059** | **5.062** | **5.256** |
| | 6.950 | 7.136 | 6.951 | 7.161 | 7.259 | 7.338 | 7.416 |
| 3 | **4.742** | **4.422** | **4.623** | **4.589** | **4.739** | **4.713** | **4.826** |
| | 6.938 | 6.967 | 6.710 | 6.799 | 6.808 | 6.832 | 6.872 |

### 4.2 Interpolated Results from Data 2

Data 2 comprises 148 sample locations over a street network of a denser grid-like configuration (elevation level: 18~34 m) (Figure 2). Using the same set of parameters, cross-validation is conducted for all sample locations. The results are summarised in Table 2 (the one with the smaller MSE value is underscored). It is clear that PL-IDW outperformed NT-IDW in most cases, as can be seen in Figure 2 through the comparison of the mean value of their $D$ index (black points indicate locations where NT-IDW yielded more error). The fact that they are on the periphery of the network suggests that the edge effect of the network (Ripley 1981) may be accountable. The discontinuity of the street network in the periphery of the study area prevents us from measuring the shortest-path distance from the sample locations to the target location; whereas for PL-IDW, all nearby sample locations can be identified.

### 4.3 Interpolated Results from Data 2 after Modification

In order to eliminate the edge effect and conduct the cross-validation of NT-IDW and PL-IDW under similar conditions, a guard area is introduced to Data 2. The data is modified by (1) adding network segments outside the study area until nearby sample locations for all target location are accounted for; and (2) adding the sample locations from the extended segments to the original sample locations. Results are shown in Figure 3, where the cross-validation is carried out with modified Data 2 (104 sample locations are added to the original 148 locations). The change in the combinations of the

nearby sample locations for NT-IDW has played a critical role in improving MSE. It shows that even though the edge effect affects both the planar and network cases, the impact is greater on network.

Table 3 summarises the results from the cross-validation of modified Data 2. There are two notable differences from Table 2. First, the overall prediction accuracy is improved for both NT-IDW and PL-IDW. Second, NT-IDW predicts more accurately than PL-IDW does in most cases.

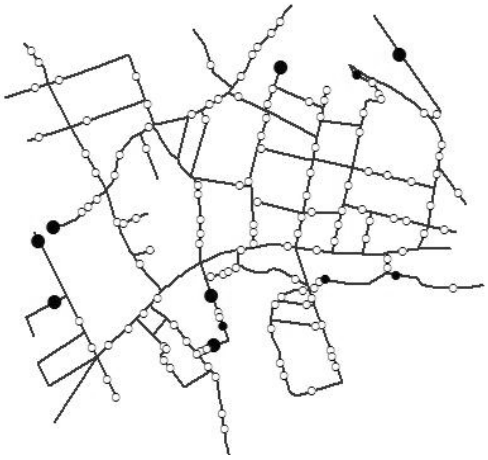


**Figure 2**. A grid-like street network with 148 sample locations. Locations at which NT-IDW outperformed PL-IDW are shown in white circles, and those at which NT-IDW under-performed are shown in black. Small circles show those with average *D* index value between 3 and 5, and large circles locations where *D*>5 (i.e. NT-IDW gave a particularly poor estimate).

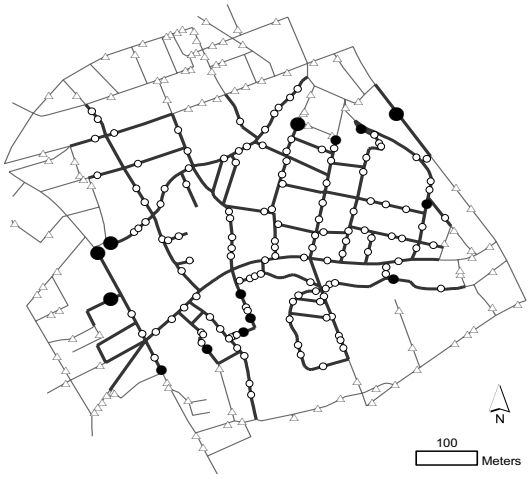



**Figure 3.** Increase in the level of prediction accuracy of NT-IDW at each location of modified Data 2. Black circles show sample locations at which the prediction accuracy showed improvement (large circles: high accuracy with average *D* index greater than 3, and small circles: average *D* index between 1 and 3).

**Table 2.** MSEs of the planar and the network interpolations of Data 2.

| λ \ s | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 0.5 | **2.345**<br>2.504 | 2.747<br>**2.530** | 3.276<br>**2.920** | 3.532<br>**3.161** | 3.870<br>**3.604** | 4.091<br>**3.655** | 4.414<br>**3.806** |
| 1 | **2.256**<br>2.394 | 2.514<br>**2.367** | 2.886<br>**2.632** | 3.055<br>**2.783** | 3.282<br>**3.103** | 3.430<br>**3.121** | 3.645<br>**3.258** |
| 1.5 | **2.233**<br>2.349 | 2.405<br>**2.296** | 2.654<br>**2.466** | 2.757<br>**2.541** | 2.898<br>**2.759** | 2.989<br>**2.741** | 3.121<br>**2.848** |
| 2 | **2.250**<br>2.347 | 2.371<br>**2.282** | 2.539<br>**2.394** | 2.600<br>**2.420** | 2.686<br>**2.567** | 2.740<br>**2.528** | 2.819<br>**2.606** |
| 2.5 | **2.288**<br>2.369 | 2.377<br>**2.300** | 2.492<br>**2.378** | 2.529<br>**2.378** | 2.581<br>**2.482** | 2.615<br>**2.437** | 2.665<br>**2.494** |
| 3 | **2.333**<br>2.403 | 2.401<br>**2.334** | 2.483<br>**2.394** | 2.506<br>**2.380** | 2.537<br>**2.456** | 2.560<br>**2.415** | 2.592<br>**2.456** |

**Table 3.** MSEs of the planar and the network interpolations of modified Data 2.

| λ \ s | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 0.5 | **1.926**<br>2.189 | **2.261**<br>2.325 | **2.366**<br>2.514 | **2.324**<br>2.629 | **2.453**<br>2.929 | **2.699**<br>2.934 | **2.744**<br>2.986 |
| 1 | **1.881**<br>2.127 | **2.097**<br>2.194 | **2.143**<br>2.305 | **2.129**<br>2.359 | **2.220**<br>2.572 | **2.369**<br>2.546 | **2.400**<br>2.618 |
| 1.5 | **1.898**<br>2.125 | **2.049**<br>2.148 | **2.048**<br>2.200 | **2.040**<br>2.199 | **2.099**<br>2.340 | **2.182**<br>2.285 | **2.192**<br>2.355 |
| 2 | **1.951**<br>2.158 | **2.065**<br>2.152 | **2.039**<br>2.173 | **2.031**<br>2.134 | **2.069**<br>2.226 | **2.113**<br>2.156 | **2.111**<br>2.216 |
| 2.5 | **2.018**<br>2.208 | **2.110**<br>2.182 | **2.074**<br>2.191 | **2.066**<br>2.129 | **2.040**<br>2.199 | **2.114**<br>2.119 | **2.108**<br>2.170 |
| 3 | **2.088**<br>2.264 | **2.165**<br>2.224 | **2.128**<br>2.230 | **2.120**<br>2.156 | **2.134**<br>2.199 | 2.149<br>**2.134** | **2.142**<br>2.175 |

## 5. Conclusion

This study extended IDW interpolation to the network space. The proposed method was compared with the standard IDW by means of cross-validation using two different datasets. Empirical analysis suggests that NT-IDW is particularly effective when applied to a network with a simpler topological structure comprising fewer segments and a large variation in the observed values, since the difference in the combinations of the nearby sample locations between the planar and the network IDWs would lead to a large disparity in their predicted values, typically in favour of NT-IDW. This difference however becomes marginal when the network takes a more regular, grid-like configuration with higher density of samples. The study also demonstrated that, even in the case of a grid-like network, a higher level of prediction accuracy can be achieved by NT-IDW if we account for the edge effect.

Although the proposed method only utilises IDW, other variations of interpolation methods may be also adopted for application in the network space.

## REFERENCES

Cressie, N., 1993. *Statistics for Spatial Data*. New York: John Wiley & Sons.
Isaaks, E.H., and R. M. Srivastava. 1989. *Applied Geostatistics*. New York: Oxford University Press.
Ripley, B.D. 1981. *Spatial Statistics*. New York: Wiley.

## Biography

Narushige Shiode is a lecturer in Spatial Analysis and GIS at School of City and Regional Planning, Cardiff University. His research interests are methods of spatial analysis and modelling, urban growth dynamics, and cartography and geo-visualisation.
Shino Shiode is a lecturer in GIScience at Department of Geography, Environment and Development Studies, Birkbeck College, University of London. Her research interests are methodology of spatial analysis with a particular focuses on crime, health and urban applications.

# Modelling Spatial Association between Temples and Land-use Change

Shihong Du[1], Luo Guo[2], Robert Haining[3]

[1]Institute of Remote Sensing and GIS, Peking University, Beijing 100871, P.R. China
Email: dshgis@hotmail.com

[2]College of Life and Environmental Science, Minzu University of China, Beijing, 100081, China

[3]Department of Geography, University of Cambridge, Downing Place, Cambridge, CB2 3EN, UK

KEYWORDS: Cross K-function; spatial association; land-use change; religious culture; spatial point pattern analysis

## 1. Introduction

Spatial associations, interactions and patterns between univariate or multivariate spatial processes play an important role in analyzing and modelling spatial data in environmental and social sciences. The association in spatial data analysis means the extent to which the statistically significant patterns about univariate or bivariate data are spatially close to each other (Haining 1990). Various methods have been developed to measure spatial associations in geo-referenced data. The point-based methods are designed to analyze spatial point process and classify the pattern of a set of points over space as random, clustered, or regular. The methods include quadrant analysis, mean nearest-neighbour distance, the distribution function of nearest-neighbour distance, Ripley's K-function (Ripley 1976), and network K-function (Okabe and Yamada 2001). Generally speaking, spatial patterns analysis depends on the spatial scale, i.e., the spatial points may be clustered at some scales, while dispersed at other scales. Among the spatial point pattern methods, Ripley's K-function can describe spatial point patterns at multiple distance scales, while other methods cannot (Dixon 2002).

The global indicators measure the overall spatial association in the entire study region. However, they hide the instability and variation of spatial associations across local areas due to spatial heterogeneity (Anselin 1995). Thus, local indicators for spatial association (LISA) have been developed to measure spatial pattern for each geographical location or individual subject. It is commonly interpreted as the local instability of spatial association and used to identify spatial outliers (Anselin 1995).

Although both global and local indicators have been widely investigated and applied, they are intended to only handle a single type of spatial data (e.g., point process, area data, categorical data, flow data, etc). In some situations, however, it is necessary to analyze the spatial association about the mixture of different types of spatial data, (e.g., lines and points, points and polygons, lines and polygons). Furthermore, these spatial associations may vary at multiple scales and over local areas.

In this study we focus on the spatial association between land-use changes and temples in the Tibetan

Autonomous Prefecture of Qinghai Province, P.R. China. The land-use changes are represented as polygons with attributes describing the types of changes, while temples are considered as points., as well as investigating the local variations of the land-use changes aggregated around the temples at different spatial scales.

To model the spatial association between land-use change and temples/villages, We extended the cross K-function (Dixon 2002; Boots and Okabe 2007) to develop global and local indicators for measuring the global spatial association between the land-use changes (polygons) and the locations of temples (points). The global indicator measures the association between points and all types of land-use change; the local indicator is investigated to explore the locally spatial clusters of the spatial association. The land-use change and temples data are collected to analyze the spatial association and its variations with space. The global indicators show that the temples are positively correlated with land-use change. The local patterns show great variations.

## 2 Global and local indicators for point and polygon patterns

To model the association, the control points and the event points, the circular neighbourhood around a control point, and the density of event points in the neighbourhood are essential components. In our study, control variable is the point pattern (such as temples and villages); the event variable is the polygon pattern representing land-use change. The indicators are presented both locally and globally. The significance tests are conducted by using the Monte Carlo simulation.

### 2.1 Global indicator

The global indicator of spatial association between points and polygons is defined as:

$$PP(r) = \frac{1}{\rho * n} * \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{Area\left(\left(P_i \bigcap T_i\right) \bigcap R_j\right)}{Area(P_i \bigcap T_i)} \quad (1)$$

where, $n$ and $m$ denote the number of points and polygons, respectively; $P_i$ is the circle centred on control point $i$, $T_i$ refers to the Thiessen polygon of point $i$, and $R_j$ is a land-use change polygon. $\rho$ is the original density of land-use change (changes per unit), $\rho = \sum_{j=1}^{m} Area(R_j) \Big/ Area(U)$, where $U$ is the study region. Symbol $\bigcap$ means the intersection of two spatial objects, function *Area*() computes the area of a spatial region.

The indicator $PP(r)$ is defined as the average ratio of the density of land-use change for each control point to the original density, thus $PP(r)$ is larger than or equal to zero. The larger it is, the more the land-use changes are congregated around control points, i.e., the stronger the points have influences on land-use change. Generally speaking, within a distance scale, the more the land-use changes there

are around the control points, the larger the indicator is. This indicator can measure the spatial association between points and polygons, i.e., in what degree the polygons of land-use change are related to the distances from control points. Furthermore, given a series of distances $r_1$, $r_2$, …, $r_l$, the $PP(r_1)$, $PP(r_2)$, …, $PP(r_l)$ reveal the variation of the relationships between points and polygons at the given distances.

### 2.2 Local indicator

The global indicator $PP(r)$ measures the spatial association between points and polygons from a global perspective. The local pattern may show great variations over different points. That is, even if the point patterns are strongly associated with polygon patterns, some points may have weak links with polygons in a local space. The local indicator can help to uncover the interesting patterns different from the global one. To compute the local indicator at distance $r$, the parts of land-use polygons inside the neighbourhood is first added, and then divided by the area of neighbourhood and original density (Equation 2).

$$PP_i(r) = \frac{1}{\rho} * \sum_{j=1}^{m} \frac{Area\left(\left(T_i \bigcap P_i\right) \bigcap R_j\right)}{Area(P_i \bigcap T_i)} \tag{2}$$

It is clear that $PP_i(r)$ is concerned with a point $i$, the neighbourhood of this point, and the parts of land-use polygons inside the neighbourhood.

The Monte Carlo method is used to simulate the random pattern of the spatial association between points and polygons, and the significances of global and local indicators are tested based on the simulation. According to the significance tests, both the local and global indicators fall into three patterns: clustered, random, and regular ones. A clustered pattern means a large rate of land-use change around points; a dispersed pattern implies a small the rate; while for a random, the pattern is not significant from the statistical perspective.

## 3 Case studies

Two case studies are conducted to analyze the interactions between temples points and polygons of land-use change. The temples are the centres of Buddhist culture and ceremonies. The first case uses the global indicator to measure the spatial association between temples and land-use change. The results show that the indicator is statistically significant when the distance is larger than or equal to 6,000 m; while it is insignificant when the distance is less than 6,000 m. The second case investigates and visualizes the variation of local patterns between temples and land-use change. In Figure 2, symbols '★', '⬣', and '▲' represent the random, regular and clustered patterns between this temple and the polygons. Among the 66 temples, there are 45 regular, 16 clustered and 5 random patterns at 1,000 m scale; 43 regular, 18 clustered and 5 random patterns at 9,000 m scale. Comparing the two

patterns at 1,000 m and 9,000 m scales, 10 local patterns have changed.

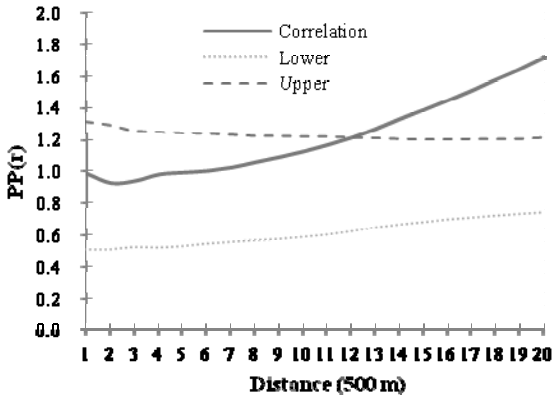



Figure 1. Global indicator between temples and land-use changes

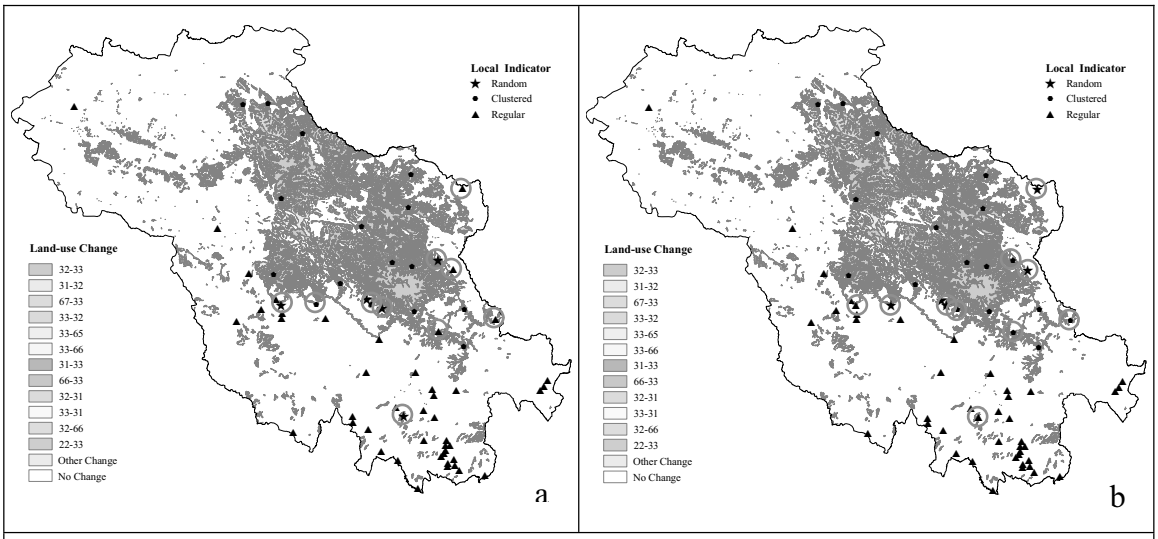



Figure 2. Local patterns between temples and land-use change at different distances: (a) *radius*=1000 meter; (b) *radius* =9000 meter.

## References

Anselin L (1995) Local indicators of spatial association *Geographical Analysis* **27(2)** 93-115.

Boots B and Okabe A (2007) Local statistical spatial analysis: inventory and prospect *International Journal of Geographical Information Science* **21(4)** 355-375.

Dixon P M (2002) Ripley's K function. In: Encyclopedia of Environmetrics, El-Shaarawi A H and Piegorsch W W (Eds.) John Wiley & Sons, Ltd, Chichester, pp 1796-1803

Okabe A Yamada I (2001) The K-function method on a network and its computational implementation *Geographical Analysis* **33(3)** 271-290

Ripley B D (1976) The second-order analysis of stationary point processes *Journal of Applied Probability* **13(2)** 255-266.

Robert Haining (1990). Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge.

**Biography**

*Shihong Du, Ph.D., associate professor in the institute of remote sensing and GIS, Peking University, P. R. China. His interesting is spatial analysis and reasoning, especially scale and uncertain issues in spatial analysis and reasoning.*

*Robert Haining, Ph.D., Professor of Human Geography, University of Cambridge, UK. He is interested in methodologies for spatial data analysis with applications in health services research, the geography of crime and economic geography.*

# A Map to Hear - Use of Sound in Enhancing the Map Use Experience

Mari Laakso and L. Tiina Sarjakoski

Finnish Geodetic Institute
Department of Geoinformatics and Cartography
Geodeetinrinne 2
FIN-02430 Masala, Finland
Tel. +358 9 2955 5130
firstname.lastname@fgi.fi
www.fgi.fi
KEYWORDS: soundscape, perceive, sonic map, use experience, visually impaired

## 1. Introduction

The way we perceive our geographical space is a combination of input from different senses. We collect many kinds of input to build up an image of the geographical space we are moving in. The image in our heads combined with other information we possess is a cognitive map which we use to orient ourselves spatially. Throughout history, maps have been an important means to provide information about places, locations and their spatial relationships. Traditional maps are 'mute', but maps can now be enlivened with sound and other multimedia in the digital era.

The possibilities of the use of sound have been recognised by many authors. Krygier (1994) presented a variety of sound forms and the wide range of possibilities to use them in geographic visualisation applications. The concept of soundscape was introduced by Schafer (1977, 1994). He defined and described the elements of a sonic environment and defined the soundscape as (p. 274): "Technically, any portion of the sonic environment regarded as a field for study. The term may refer to actual environments, or to be abstract constructions such as musical compositions and tape montages, particularly when considered as an environment."

The potential of soundscapes in cartography has been discussed by Théberge (2005). He recognised the figure-ground relationship in environmental sound and its possibilities in maps. Théberge continued with the concept of cypercartography originated from Taylor (1997). The use of soundscapes in enriching our multisensorial reading of space is discussed by Caquard et al. (2008) as well as the role of sound in cartography more widely. Brauen and Taylor (2007) discussed the motivation for introducing multisensory information into mapping projects. They presented a framework for the incorporation of sound into visual maps with a realised example of an atlas project.

Rice et al. (2005) faced the design of map interfaces in a context of visually impaired users with a project 'Haptic Soundscapes'. Sarjakoski et al. (2009b) discussed about the importance of sound maps for use experience. An example of using soundscapes on a map comes from the city of Tampere, Finland, where *Aamulehti*, the main local newspaper, constructed a soundscape of Tampere for the Web.

With digital maps, not only visual but other means of expression are possible. Hearing comes in the second place in our senses after seeing, so attaching sound to a map is worth studying. As stated by Brauen and Taylor (2007) cartography cannot afford to continue to undervalue the use of sound as well as the other non-visual senses.

The motivation of this research, carried out as part of two current research projects, is to

communicate spatial information even to those who are visually impaired and, in general, provide map users with a more profound use experience. Perceptualization through multiple channels is one of the key principles identified in the requirements for the HaptiMap –project, in which LBSs accessible for all are being developed (Magnusson et al, 2009). In the MenoMaps project, the enrichment of map use experience is given special attention in developing a multi-publishing web map service (Sarjakoski, 2009).

In the following, two examples in Sections 2 and 3 on sound map implementations are described and discussed. The applications cover a hiking use case with the aim to provide tools for a user to plan a hike in advance and also obtain information about the environment.

## 2. Design and implementation of the soundscape map

We have started the experiments with embedding sound landscapes into hiking maps. The motivation of this research is to communicate spatial information and provide map users with a profounder experience (Sarjakoski et al., 2009a). Soundscapes can serve many kinds of users in perceptualizing spatial information – for example, in indoor use when planning a hike.

In this study, the recorded soundscapes could be compared with photographs. They are shots taken at a certain point at a certain moment. Hence, they are audiorealistic and include the sonic environment of an instant. Like photos, audio records may vary significantly depending on the moments of the shot taken. For instance in springtime, by comparison to a snowy winter day, there is a chorus of birds singing and creeks full of gurgling water. The recordings were made with an Olympus linear PCM recorder LS-10.

The implementation of the soundscape map was carried out with Adobe Flash CS3 software. On a web hiking map, play button symbols were added to locations where soundscapes were recorded in the field. When pressing the play button, an MP3 sound file starts playing. The field recorded soundscape of the location is in the file. When the button is pressed down, it will turn into a stop button and the user can stop the sound by pressing it again. In addition to the audio response of the button, a small text box also appears, containing some additional information about the recording such as time and place (Figure 1).

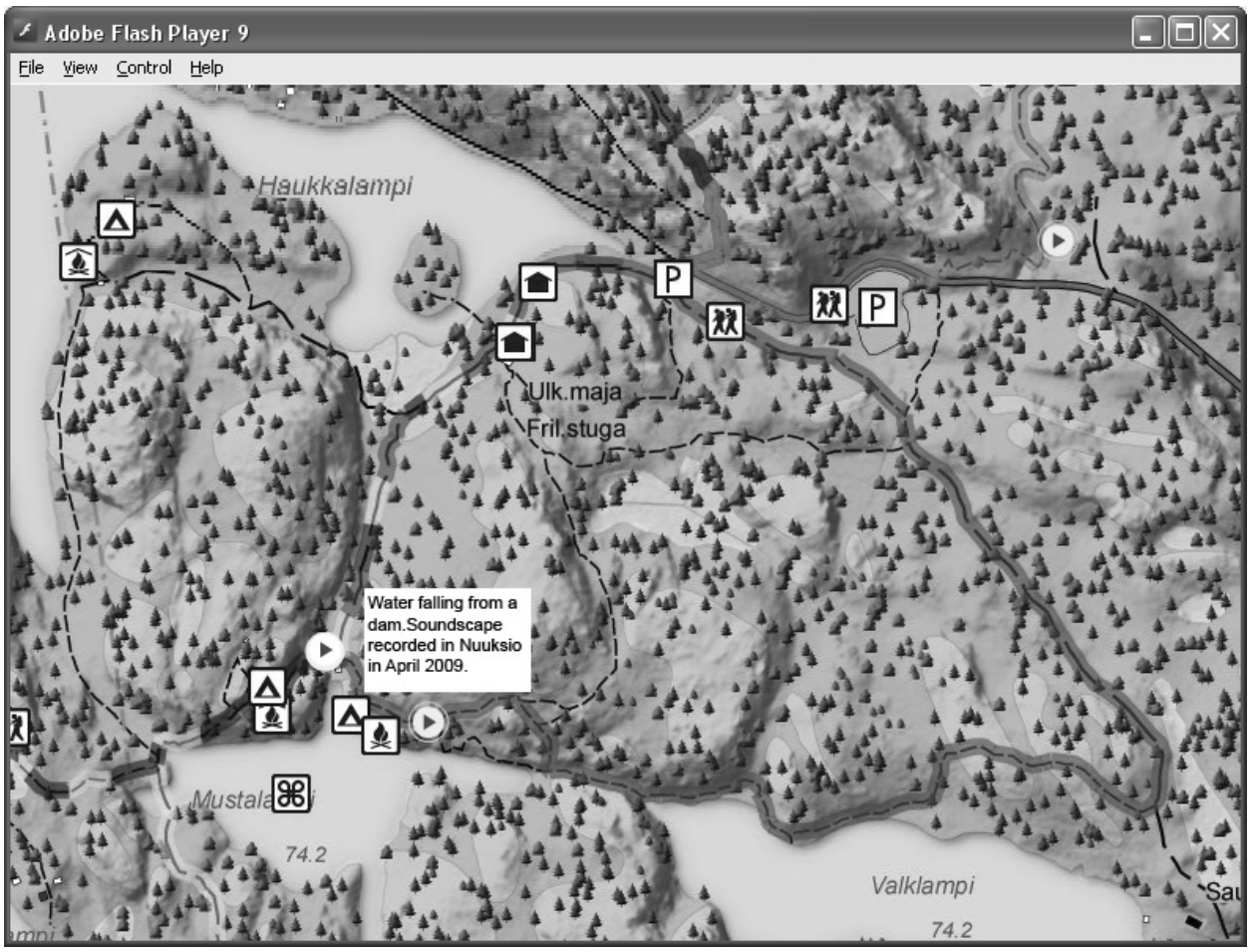



**Figure 1.** The soundscape map over Nuuksio, integrated with sound buttons. The background map originates from the MenoMaps project (Oksanen et al., 2010).

When embedding a sound dimension into the map, a reflection of the total cognitive load caused to the map user should be considered. When displaying a map on a digital device, the possibility exists for interaction with the user. The user is more or less able to choose what s/he wants to see and hear. Since sounds can in certain situations be very irritating, there should always be at minimum the possibility to mute or stop the action. Especially with sounds mediating fine, polyphonic ambience, the listening situation also plays a big role. Moreover, cultural differences may cause various unforeseeable and even faulty interpretations.

### 3. Design and implementation of the sonic map for visually impaired

Traditionally maps have been purely visual, thereby giving either no help or very little to blind or otherwise visually impaired people. For people who cannot rely on their vision, the importance of other senses is emphatic. People who are visually impaired but not totally blind could benefit from a map with embedded sound effects, primarily designed for the visually impaired (Magnusson et al., 2009; Sarjakoski et al, 2009a). For weak-sighted users, an overall map-design with reduced and simplified visual cues is essential. Compared with traditional map layouts, the visual appearance has to be radically generalised, the colours need to be enhanced and contrast must be increased. After the visual simplification, the sonification of the map takes place. The visual map textures get analogous sound textures when all the map elements are attached with characteristic or representative sounds.

Visually impaired people may, with certain limits, utilise visual maps, especially when displayed on

large lighted screens; however some confusion is possible. For example, it is essential information for the user whether the line on the map represents a path or a ditch. A quick response to this uncertainty may be provided for the user by sonic information embedded in a map object and captured through a mouse-over action. Also, the symbols presenting different kinds of spatial objects on the map may be either simply read out or integrated with a symbolic sound or a spoken guiding description. Similarly, all the written text on the map can be read out through a mouse-over action.

In this experiment, an intensive map generalisation of the base map was carried out first. The base map itself was created in the MenoMaps –project (
Oksanen et al., 2010;Sarjakoski, 2009). The walking routes were emphasized with thicker lines having generalised simplified forms, and the colours were given special attention. Water areas were also enhanced. The detailed information of the map, such as contour lines, was faded out into the background by using transparent colour. We did not want to delete it totally since in our opinion it may serve the accompanying person. After the visual generalisation, the sonification was added.

This realization was completed using Adobe Flash CS3 software. The various areas and separate objects of the map were integrated into invisible buttons. Thereafter, the sound files were attached to the buttons, and after that the sounds could be listened through a mouse-over function. The user may explore the map with mouse (or other pointer) and in every location hear a sound related to the object, Figure 2. The sounds that were used in sonification were:

- for the forest areas: the on-site recorded sounds of forest (singing birds, some wind)
- for the water: the gurgle of a brook (from one on-site brook, used in all)
- for the paths: walking footsteps on a dirt path
- for the roads: the sounds from cars
- the map symbols and the text are read out

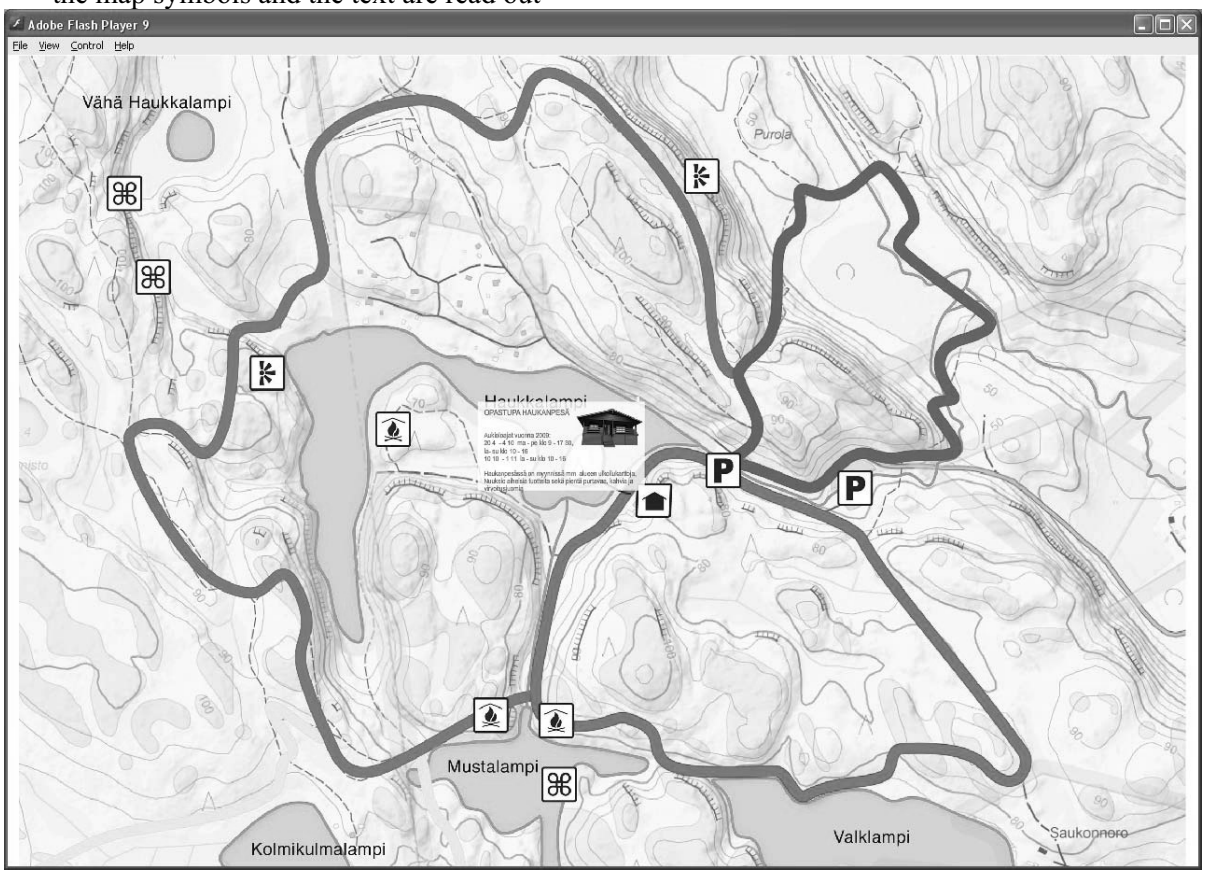



**Figure 2.** The sonic map from the Nuuksio test environment. The sounds of various map objects can be explored through mouse-over actions. The background map originates from the MenoMaps project (Oksanen et al., 2010).

## 4. Summary and conclusions

The primary use of the presented soundscape map applications is to serve stationary users when planning a hike or users who are unable to visit the actual place. An example of such a user group is people who have restricted locomotion ability. With a sonified map, visually impaired users can familiarise themselves with the area in advance and find sonic landmarks in order to obtain help in recognising places when later visiting, e.g. a national park. To the users unable to go on a hike at all, the real world sounds from the forest embedded in the map mediate the true ambiance and thus provide a sort of accessibility into the nature. Sound in maps can serve all kinds of map users by providing essential and additional information, depending on the current situation. The possibilities of sonic maps including real world sounds and other aural means deserve further research. Our next step will be user testing with various user groups.

## 5. Acknowledgements

This survey is a part of research in two projects. The European Commission –supported HaptiMap – project (FP7-ICT-224675) is coordinated by Lund University's Department of Design Sciences (www.haptimap.org). The MenoMaps -project, funded by Tekes (the Finnish Funding Agency for Technology and Innovation), is a joint venture of the Finnish Geodetic Institute, (Department of Geoinformatics and Cartography) and The Department of Design of the Aalto University School of Art and Design.

## References

Brauen, G. and D.R.F. Taylor (2007). A cybercartographic framework for audible mapping. Geomatica, 61(2), pp. 19-27.

Caquard, S., G. Brauen, B. Wright and P.Jasen (2008). Designing sound in cypercartography: From structured cinematic narratives to unpredictable sound/image interactions. International Journal of Geographical Information Science, vol.22, nos 11-12, pp. 1219-1245.

Krygier, J.B. (1994). Sound and geographic visualization. In: A.M. MacEachren, A.M. and Taylor, D.R.F. (eds.), Visualization in Modern Cartography, Pergamon New York, pp. 149-166.

Magnusson, C., Brewster, S., Sarjakoski, T., Roselier, S., Sarjakoski, L.T. and K. Tollmar, 2009. Exploring Future Challenges for Haptic, Audio and Visual Interfaces for Mobile Maps and Location Based Service, LocWeb 09 Workshop, CHI 2009, April 4 – April 9, 2009, Boston, MA, USA.

Oksanen, J., Schwarzbach, F., Sarjakoski, L.T., and T. Sarjakoski (submitted). New design for the Finnish Nuuksio National Park maps – case MenoMaps. 7th ICA Mountain Cartography Workshop, 1-5 Sep 2010, Borsa-Maramures, Romania.

Rice, M., Jacobson, D., Golledge, R.G. and D. Jones (2005) Design Considerations for Haptic and Auditory Map Interfaces, Cartography and Geographic Information Science, vol. 32(4), pp. 381-391.

Sarjakoski, L. T., 2009. Jokapaikan spatiaalinen vuorovaikutus, (Eng. Ubiquitous Spatial Interaction). Positio, (4/2009): 6-8, (in Finnish).

Sarjakoski, L. T., Ylirisku, S., Flink, H.-M. and SWeckman (2009a). Explorative User Study Approach for LBS Innovation for Hikers. Proceedings of the 24th International Cartographic Conference, 15-21 Nov, 2009, Santiago, Chile, Theme 17: Users.

Sarjakoski, L. T., Sarjakoski, T., Koskinen, I., S. Ylirisku (2009b) The Role of Augmented Elements to Support Aesthetic and Entertaining Aspects of Interactive Maps on the Web and Mobile Phones, In: Cartwright, W, Gartner, G., Lehn, A., (eds.), Art and Cartography, Springer Berlin Heidelberg

New York, pp. 109–124.

Schafer, R.M. (1994, 1st ed. 1977). The Soundscape: Our Sonic Environment and the Tuning of the World. Rochester: Destiny Books.

Tampereen äänimaisemakartta (2008) Online: http://www.aamulehti.fi/aanimaisema/ (08/01/10).

Taylor, D.R.F. (1997) "Maps and Mapping in the information era", Keynote address to the 18th ICA Conference, Proceedings, Vol. 1, Swedish Cartographic Society, Stockholm, Sweden, pp.1-10.

Théberge, P. (2005). Sound Maps: Music and Sound in Cybercartography. In: Taylor, D.R.F. (ed.). Cybercartography: Theory and Practice, Elsevier Amsterdam, pp. 389-410.

**Biography**

*M.Sc. Mari Laakso is a senior research scientist at the Finnish Geodetic Institute in the Department of Geoinformatics and Cartography. Her interests are in multimedia and augmented cartography. She has graduated from the Helsinki University of Technology with a Master of Science degree in Land Surveying.*

*Docent, Lea Tiina Sarjakoski acts as a chief research scientist at the Dept. of Geoinformatics and Cartography. She obtained the degree of D.Sc. (Tech.) at Helsinki University of Technology in 1997. She has published in the field of knowledge-based methods for GI, usability of mobile topographic maps, visualisation and generalisation for network processing and mobile map services.*

# Using Sound to Represent Uncertainty in Address Locations

Nick Bearman[1] and Andrew Lovett[1]

[1]School of Environmental Science, University of East Anglia, Norwich, NR4 7TJ
Tel: +44 (0)1603 591346, www.nickbearman.me.uk
Email: n.bearman@uea.ac.uk, a.lovett@uea.ac.uk

KEYWORDS: positional accuracy, sound, sonification, tone, ArcGIS

## 1. Introduction

All spatial data have positional uncertainty, which sometimes users ignore. There are a number of different ways of representing these data visually, but they have limitations which can be addressed by utilising non-visual methods. This study uses piano notes to represent spatial uncertainty in Ordnance Survey Address Layer 2 (AL2) data, and was evaluated by 49 spatial data users using computer based evaluations and discussion sessions.

## 2. Literature Review

The AL2 data set has different status flags within it indicating various factors including positional quality accuracy. This is important because greater uncertainty may mean addresses are shown up to several km away from their true location and ignoring this could have serious consequences (e.g. for routing applications in emergency services usage). Despite this, a number of interviews with Ordnance Survey Account Managers and Pre and Post Sales staff suggested that often status flags are not properly considered when the data is used by external organisations.

The representation of uncertainty has significant coverage within the literature and visual methods such as colour, blurring or multiple maps are generally effective (Appleton et al., 2004; Ehlschlaeger et al., 1997). However these methods can obscure underlying data or limit the amount of information shown. The use of other senses has been explored to address these limitations. Haptic (touch) maps are being developed, but they require specialised hardware and training to use (Golledge et al., 2005). Sound is also being researched and the hardware required (sound card and headphones/speakers) is readily available, but user training is required (Pauletto & Hunt, 2009).

One way of showing uncertainty information more effectively is to use sound in combination with vision, and this has been addressed from a theoretical and practical point of view (Krygier, 1994; Fisher, 1994). MacVeigh & Jacobson (2007) developed a prototype which sonified three different land uses (sea, land and harbour). They found participants understood the map quickly and that the sound enhanced their experience of the map. They suggested that an extension to an industry standard GIS (e.g. ESRI's ArcGIS) could be created which would use sound to represent spatial data. This integration into a commercial application would enable greater use and easier evaluation. Few of these ideas have gone beyond the proof of concept stage or had significant user testing. This field of research is still in an early stage of development and comparison is often difficult due to different terminologies and research frameworks (Frauenberger & Stockman, 2009).

## 3. Methods

An ArcScript (custom extension to ArcGIS) was created to allow evaluation of both visual and sonic techniques to represent positional uncertainty with Ordnance Survey MasterMap Topography and AL2 data from Norwich. The AL2 Positional Quality Accuracy status flag (PQA) was linked with the

Topography data using the associated TOIDs (topographic identifiers).

The AL2 PQA status flag values are: Surveyed (most accurate), Approximate, Postcode Unit Mean, Estimate and Postcode Sector Mean (least accurate). In a pilot study, participants said that five sounds were too many to understand effectively, as they were unable to discriminate between the different piano notes and found it difficult to relate the notes to the status flag values. The five categories were reduced to three by recoding Postcode Sector Mean to Estimate and Approximate to Postcode Unit Mean. The data presented to participants was altered to allow different proportions of Surveyed values (see below). One section of the evaluation required a second dataset to be presented; this was fabricated "Council Tax bands" information.

The three categories were represented using piano notes because the scale on a piano is very easy to visualise (i.e. participants can easily understand the difference between a 'high' and 'low' note) and the majority of people are familiar with this instrument. The notes chosen were $E_5$, $G_4$ & $C_4$, which were based on the CEG triad split over two octaves. A triad was chosen because triads are sets of notes which sound harmonious together (Burrus, 2009) and CEG was the favoured option in the pilot study. The highest note ($E_5$) represented the highest level of accuracy and the lowest note ($C_4$) represented the lowest. These were played as the participant moved the mouse over the buildings, allowing them to either query a specific building or scan an area of data to get an overall view. A legend was provided, to allow the participant to link specific notes to specific values. The data set was shown to the participant using four different presentation methods (see Table 1).

**Table 1.** The four different presentation methods (in the order they were shown to the participants) and which data were shown visually or sonically. The topography layer was always shown visually.

| **Presentation Method** | **Visual Data** | **Sonic Data** |
|---|---|---|
| Sonic only *(see Figure 1)* | Topography outlines only | AL2 Positional Accuracy |
| Visual only *(see Figure 2)* | AL2 Positional Accuracy | None |
| Visual and Sonic representing the same variable (VS Same) | AL2 Positional Accuracy | AL2 Positional Accuracy |
| Visual and Sonic representing different variables (VS Different) | Council Tax bands | AL2 Positional Accuracy |

For each presentation method, the participant was asked to identify the proportion of Surveyed values, from options of 25%, 50% or 75%. The data were randomly assigned one of these values using a stratified random method for each presentation method. Other background questions were also asked, which allowed analysis depending on musical experience, learning preference and a number of other variables. This task was chosen because it combined a simple principle (i.e. what proportion of the values are Surveyed) with the need to utilise sound in a way that visual representations are often employed.

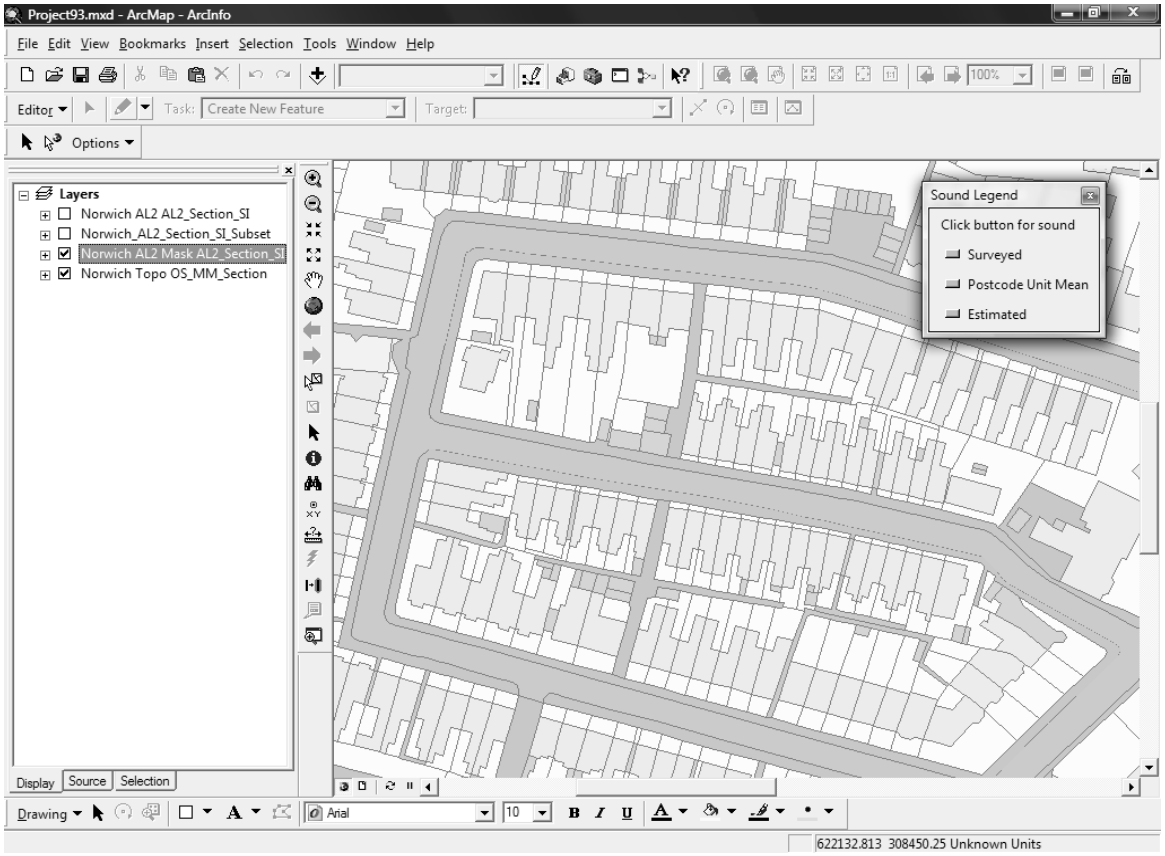



**Figure 1.** Screen shot of ArcGIS showing the data represented sonically. Note the topography layer shown visually, and the sound legend in the top right-hand corner.

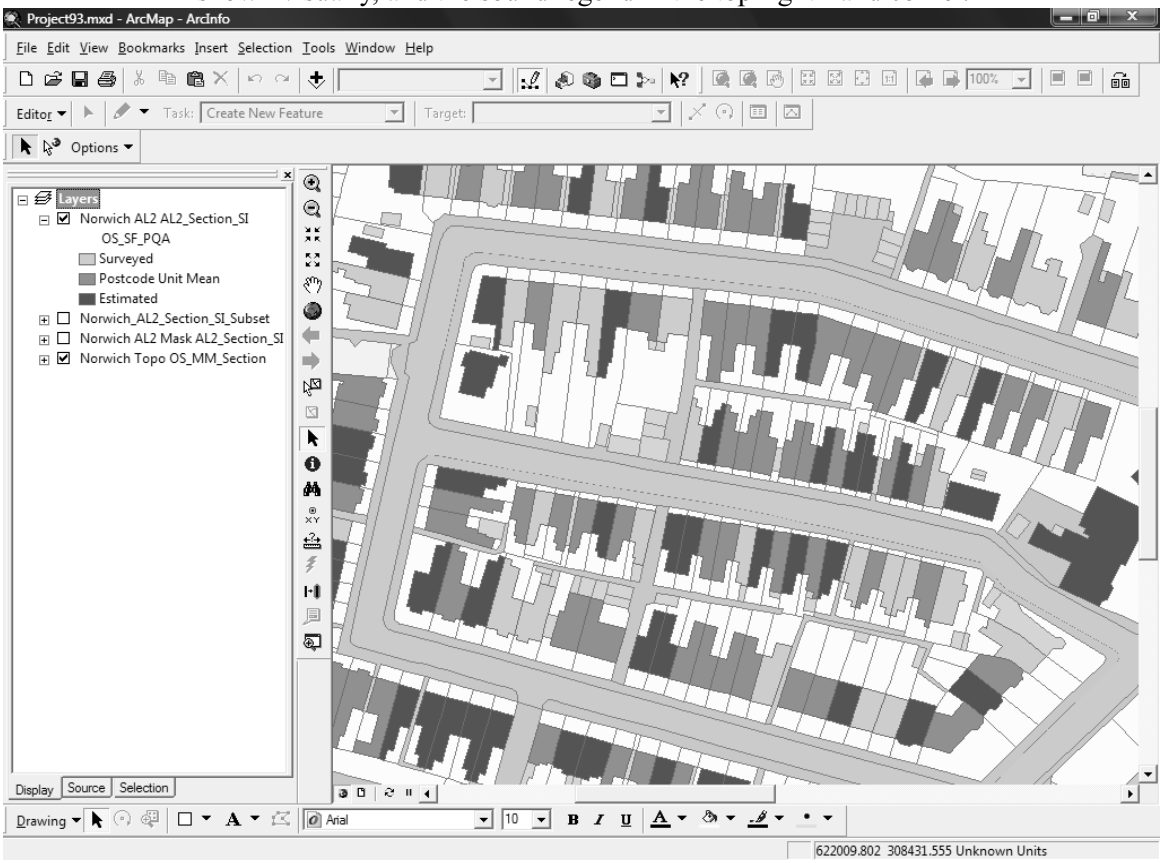


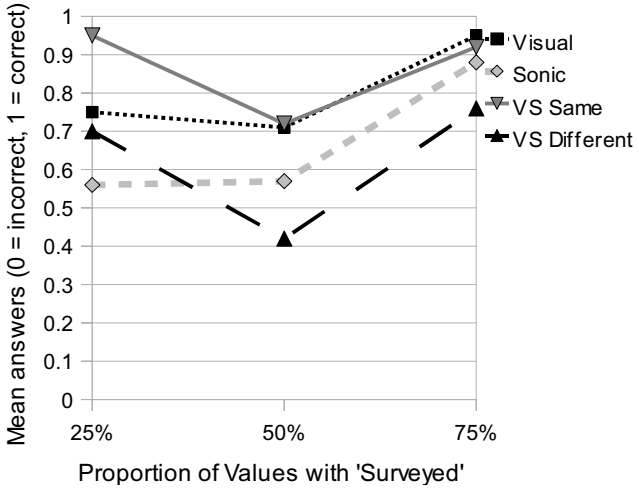

**Figure 2.** Screen shot of ArcGIS showing the data represented visually. Note the AL2 data shown visually, and the legend on the top left-hand side. The topography layer is also visible.

A total of 49 participants completed the assessment, consisting of 19 from Ordnance Survey, 23 from UEA and 7 from Local Authorities. All the participants had at least a basic knowledge of GIS, spatial data and ArcGIS and used these on a regular basis, although experience with AL2 varied. Headphones were used to provide auditory stimuli, with adjustable volume. The evaluation took place in groups of three to six, and was followed by a facilitated discussion for around 20 to 30 minutes which covered the participants' views and feelings and the potential uses of this technique.

## 4. Results & Discussion

Nearly all participants (46 out of 49) identified the correct proportion of Surveyed values for 3 or more of the presentation methods. Figure 3 shows how the mean answer for participants (correct = 1, not correct = 0) varies between proportion and presentation method, which are the two main influencing factors. The general trend was for more correct answers with 25% and 75% data proportions and less with 50% proportion. The exception to this was the Sonic presentation method, which performed as badly with 25% as it did with 50%. Visual and VS Same (see Table 1 for definitions) performed reasonably well, while Sonic and VS Different had lower correct frequencies. This could occur because the nature of the program makes it very easy to find the sounds when they are common, but difficult when they are sparse. This would impact Sonic and VS Different and potentially result in participants over estimating the proportion. Both proportion ($p < 0.005$) and presentation method ($p < 0.05$) had a significant influence on whether a participant identified the correct proportion.

**Figure 3.** Answers for all participants, split by presentation method and proportion




A higher knowledge of the data set being sonified increased the likelihood of the participants choosing the correct proportion. However, this trend was not especially strong and although including it in a logistic regression model with proportion and presentation method improved the model (see Table 2) the addition was not statistically significant.

**Table 2.** Factors added to the Logistic Regression Model and their impact

| Factors added to Model | -2 Log Likelihood | Cox & Snell $R^2$ |
|---|---|---|
| Proportion | 182.01 | 0.043 |
| Presentation Method | 169.579 | 0.11 |
| Address Knowledge | 167.319 | 0.116 |

The free text answers showed that some participants found the sonification very useful and that it added a large amount to the interpretation of the data, while others said the sound was very difficult to

understand and when combined with vision, distracted them from the visual interpretation. Sonic and VS Different were considered harder to use than Visual and VS Same. Sonic only had a very low success rate and seems unlikely to ever be successful with this type of interface.

The discussion sessions after each evaluation session provided further qualitative information and gave participants a chance to suggest changes and improvements to the technique. Preferences for the types of sounds used were subjective, and are likely to vary depending on the data set and the analysis taking place. A wider range of audio clips coupled with user choice could allow easier differentiation of sounds and potential for representation of a larger number of variables. Possibilities include different piano notes, different instruments, or completely different sounds, such as environmental or animal sounds. Colour-blind users were highlighted as a group who might find this sonification useful, however a larger sample size is required to effectively evaluate this.

The task chosen may limit the wider applicability of the results but there are few existing evaluations in this area so there is very little comparative data. The task needed to be easy enough to ensure that some of the participants managed to answer most/all of the questions correctly but not too difficult so that it would result in exclusively incorrect answers. Possible future options include more complex tasks (such as clustering exercises) and more comparisons of different presentation methods, utilising both sound and vision.

This research has highlighted specific characteristics that influence the ability of users to interpret sound to make proportion judgements. The proportion of the data the the user is interested in and the presentation method are the two factors that have the most impact on whether a person will be able to understand the proportion correctly. Knowledge of the data set being sonified also appears to have some impact, but this is not so clearly apparent with these results. These issues will be explored in future research when the author's PhD research evaluates the use of sound to represent uncertainty in UK climate scenario data.

## 5. Acknowledgements

This study has been conducted as part of an ESRC/NERC PhD Studentship. The authors would also like to thank all the participants in the study; the Ordnance Survey for their additional financial support, particularly Jenny Harding and Glen Hart for their advice, cooperation with data and assistance with arranging the evaluation sessions; and Katy Appleton for providing valuable comments while writing this extended abstract.

## References

Appleton, K., Lovett, A., Dockerty, T. & Sünnenberg, G. (2004) Representing Uncertainty in Visualisations of Future Landscapes. In: *Proceedings of the XXth ISPRS Congress*.

Burrus, C. (2009) There's Math behind the Music! [Internet]. Available from: <http://www.charlieburrus.com/MathInMusic/Index.htm> [Accessed 17 February 2010].

Ehlschlaeger, C.R., Shortridge, A.M. & Goodchild, M.F. (1997) Visualizing spatial data uncertainty using animation. *Computers & Geosciences*, 23 (4), pp.387-395.

Fisher, P.F. (1994) Hearing the Reliability in Classified Remotely Sensed Images. *Cartography and Geographic Information Systems*, 21 (1), pp.31-36.

Frauenberger, C. & Stockman, T. (2009) Auditory display design-An investigation of a design pattern approach. *International Journal of Human-Computer Studies*, 67 (11), pp.907-922.

Golledge, R.G., Rice, M. & Jacobson, R.D. (2005) A Commentary on the Use of Touch for Accessing On-Screen Spatial Representations: The Process of Experiencing Haptic Maps and Graphics. *The Professional Geographer*, 57 (3), pp.339-349.

Krygier, J.B. (1994) Sound and Geographic Visualization. In: *Visualization in Modern Cartography.* Oxford, UK, Elsevier Science, pp.149-166.

MacVeigh, R. & Jacobson, R.D. (2007) Increasing the dimensionality of a Geographic Information System (GIS) Using Auditory Display. In: *Proceedings of the 13th International Conference on Auditory Display.* Available from: <http://www.music.mcgill.ca/icad2007/proceedings.php> [Accessed 25 May 2008].

Pauletto, S. & Hunt, A. (2009) Interactive sonification of complex data. *International Journal of Human-Computer Studies*, 67 (11), pp.923-933.

## Biography

*Nick Bearman completed his MSc GIS at University of Leicester in 2008 and is currently studying for a PhD at UEA in Environmental Science, researching different methods of representing uncertainty in a variety of spatial data environments.*

# vizLib: Using The Seven Stages of Visualization to Explore Population Trends and Processes in Local Authority Research

Robert Radburn, Jason Dykes, Jo Wood

giCentre, School of Informatics, City University London, EC1V 0HB
Telephone: +44 (0)20 7040 0212, Fax: +44 (0)20 7040 8584
E: {sbbd476 | jad7 | jwo} @soi.city.ac.uk, W: http://www.soi.city.ac.uk/~ {sbbd476 | jad7 | jwo}

KEYWORDS: libraries, local authority, exploratory data analysis, visualization, *Processing*

## 1. Introduction

We use data visualization to explore patterns in a large data set of library loans maintained by Leicestershire County Council (LCC). This work has resulted in hypotheses and insights that may influence policy. Such an approach is increasingly accessible to local authority researchers through high-level languages and toolkits including *Processing* (Fry and Reas, 2007), *Prefuse* (Heer et al., 2005) and *ProtoVis* (Heer & Bostock, 2009). Fry (2008) proposes a seven-stage process for manipulating and making sense of data from its acquisition through its visualization and ultimately to its interpretation. His model (Figure 1) is not rigid, but draws attention to the interdependencies between the various elements of the visual data analysis process. It may be useful as a framework for data visualization in local authorities.

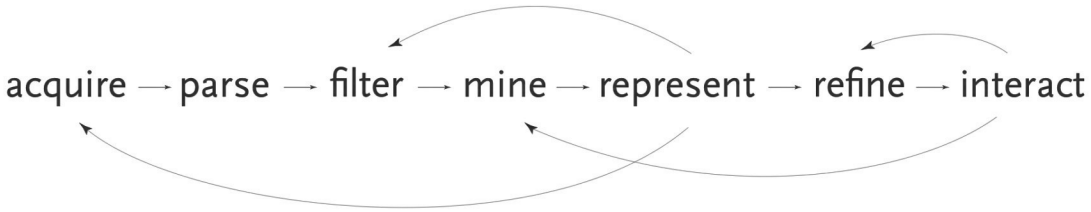



**Figure 1:** The Seven Stages of Visualization (Fry, 2004)

The model structured our analysis and we use it here to describe data visualization with Fry's open set of software tools '*Processing*' (Fry, 2008) [1]. This 'software sketchbook' is being used in a variety of scientific domains to rapidly explore data with flexible interactive visualization prototypes (e.g. Meyer et al., 2009; Slingsby et al., 2009; Wood et al., *in press*).

Before embarking on the seven stages of visualization, questions must be considered that are answerable with the data in hand: "*the more specific you can make your question, the more specific and clear the visual result will be*" (Fry, 2008.) We used the Leicestershire Library Services (LLS) TALIS database to address three questions pertinent to LLS:

- How does performance vary across the 54 libraries in Leicestershire?
- In which areas are the 'best' customers living?
- Can the area you live in contribute to predictions of usage?

[1] http://processing.org

## 2. The Seven Stages of Data Visualization

**Stage 1. ACQUIRE – Obtain the Data**
The TALIS database contains a rich spatio-temporal record of every item loaned from a Leicestershire library. Data were obtained for 435,000 active library users over a two-year period.

**Stage 2. PARSE – Structure the Data**
Fry's approach diverts effort from database design into analytical sketching (Fry, 2008). 'Flat' text files were produced from TALIS with a consistent structure for visual prototyping.

**Stage 3. FILTER – Identify the Data of Interest**
Data quality checks where undertaken and personal data removed from records. Population weighted centroids and OAC (Vickers et al., 2005; Vickers and Rees, 2007) codes were added to output areas (OAs) - the geography used to consider the home locations of library members.

**Stage 4. MINE – Methods to Uncover Patterns in the Data**
The 'recency / frequency' marketing technique used in LCC to segment records into quintiles was applied to structure the data (Radburn et al., 2007). Members of each library were categorized into one of twenty-five recency / frequency (RF) combinations. Signed chi statistics were calculated for individual libraries and OAs in line with the research questions to relate observed numbers with values expected according to regional population weighted averages.

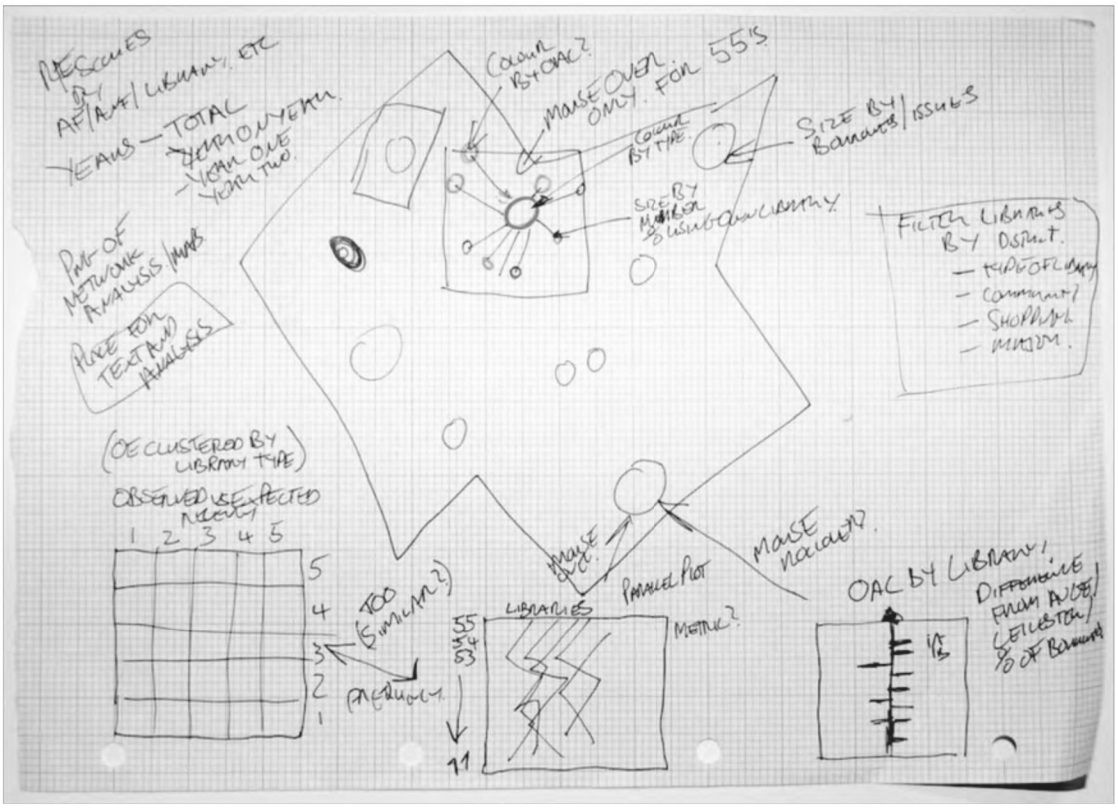



**Figure 2:** Sketch produced with LLS marketing manager outlining ideas on how the data could be depicted and interrogated. Note RF matrices (bottom left).

**Stage 5. REPRESENT – Visual Model for the Data**

*Processing* is designed for small-scale visualization that can be rapidly modified as the process of visual enquiry progresses. It encourages an exploratory approach to data visualization as code that links graphical methods can be quickly configured and applied to text files that have been parsed, filtered and mined (Radburn et al., 2009). Understanding of the context in which LLS use their data was developed through sketching with paper and pencil (Figure 2). The ideas generated were rapidly incorporated into interactive *Processing* 'data sketches' that loaded flat files generated as a result of the parsing, filtering and mining of the TALIS data (Figures 3-7).

## 3. Findings

Various graphical techniques were developed in *Processing* to depict the filtered and mined data and combined with dynamic behaviours to interactively explore the three research questions:

**(i) In which areas are the best customers living?**

**Spider plots** – LLS wanted to link library and customer (see Figure 2, top centre). Lines linking libraries with the home locations of particular groups of their users give an indication of how the different facilities compete. For example, many of the most frequent and recent users of Market Harborough library come from nearby villages containing libraries (Figure 3).

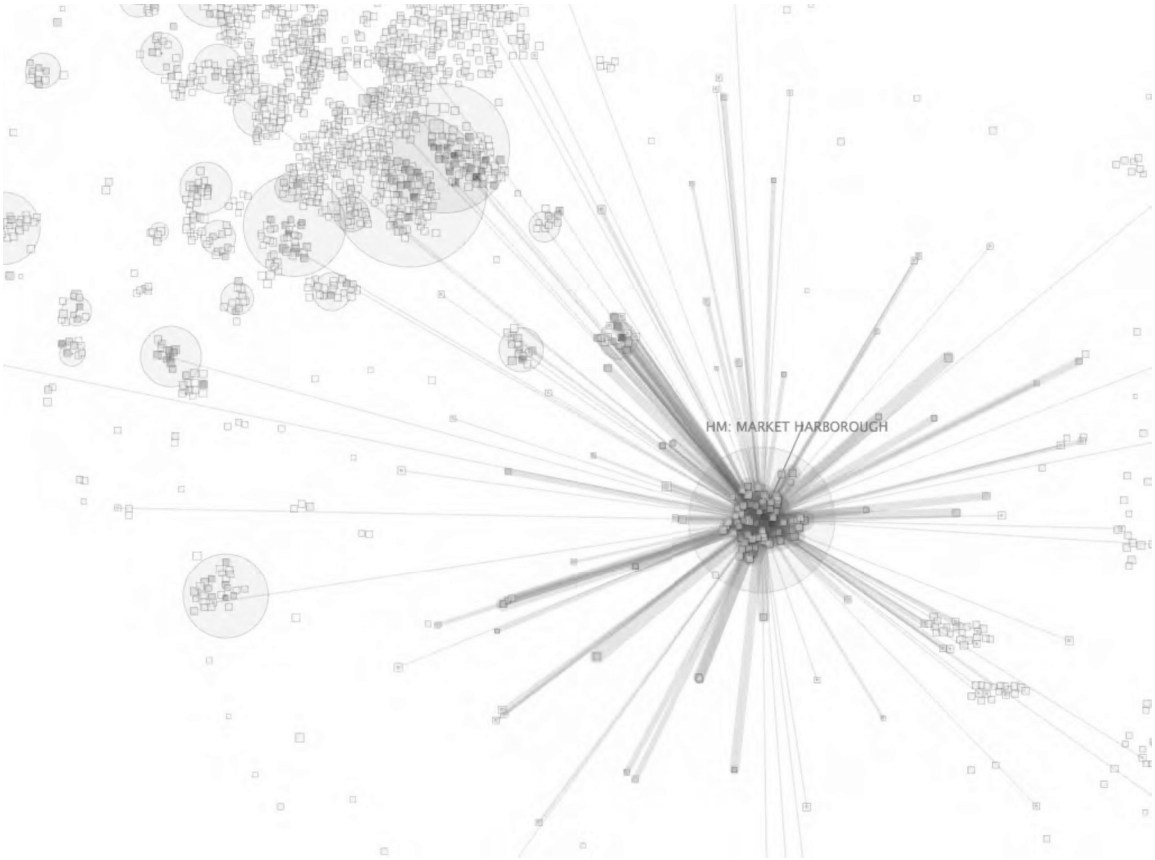



**Figure 3:** Locations of Market Harborough library 'best users' and competing village libraries. OAs are coloured with diverging 'RdBu' scheme showing variation from expected membership levels given the regional proportions (red higher than expected, blue lower). Library locations are shown using British National Grid with symbol sizes representing number of registered users.

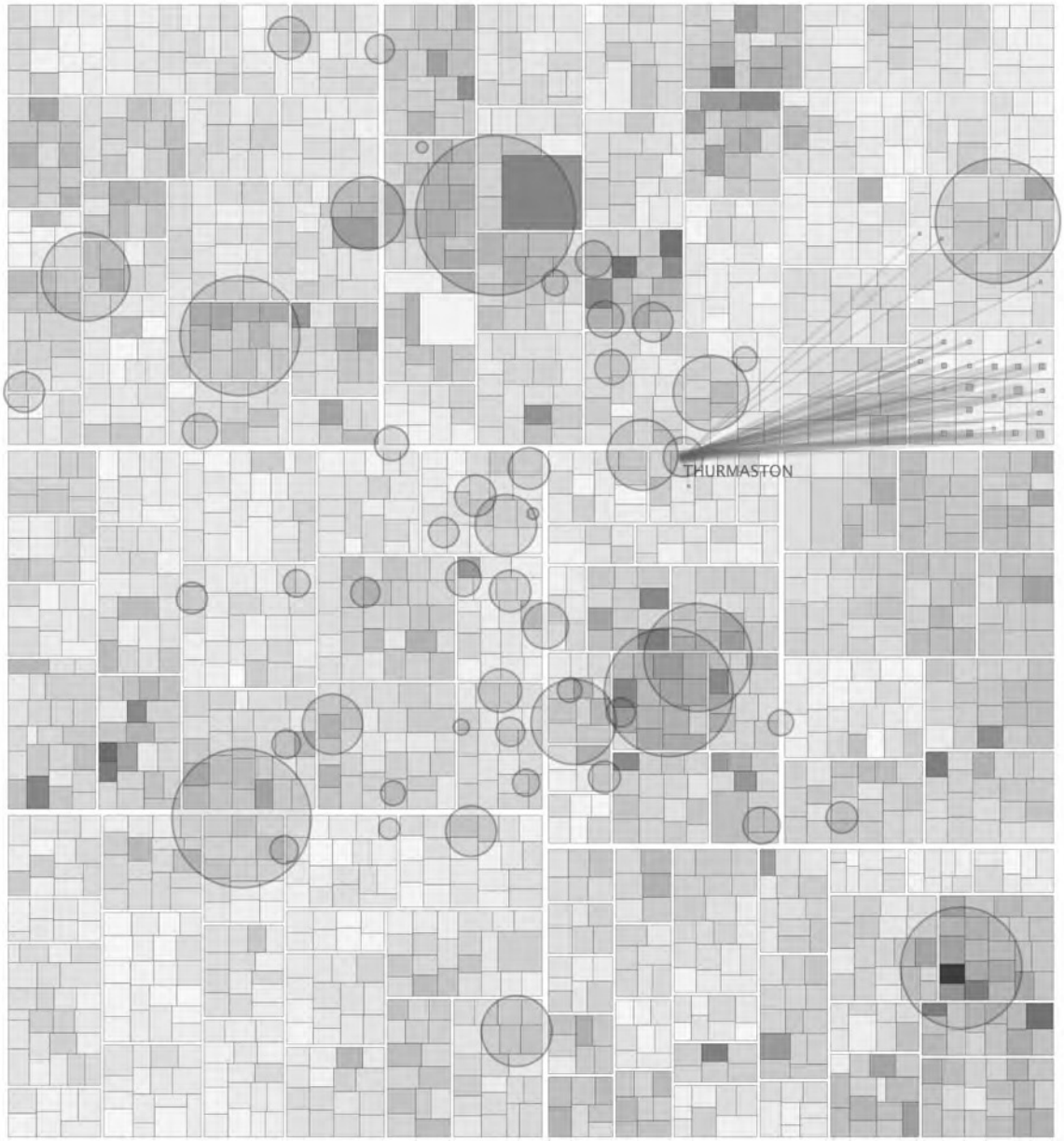



**Figure 4:** Spatial treemap of Leicestershire OAs with diverging 'RdBu' scheme showing variation from expected membership (red higher than expected, blue lower). Library locations are shown using British National Grid with symbol sizes representing number of registered users. Users of Thurmaston library are highlighted.

**(ii) Can the area you live in contribute to predictions of usage?**

**Spatial Treemaps** – These non-occluding space-filling layouts that preserve aspects of underlying geography (Wood and Dykes, 2008) may reveal subtle patterns in the data. Local variations in library performance became evident through these data rich views. These include lower numbers of members than expected in Thurmaston and subsequent consideration of possible explanations, more sophisticated modelling and local service provision (Figure 4).

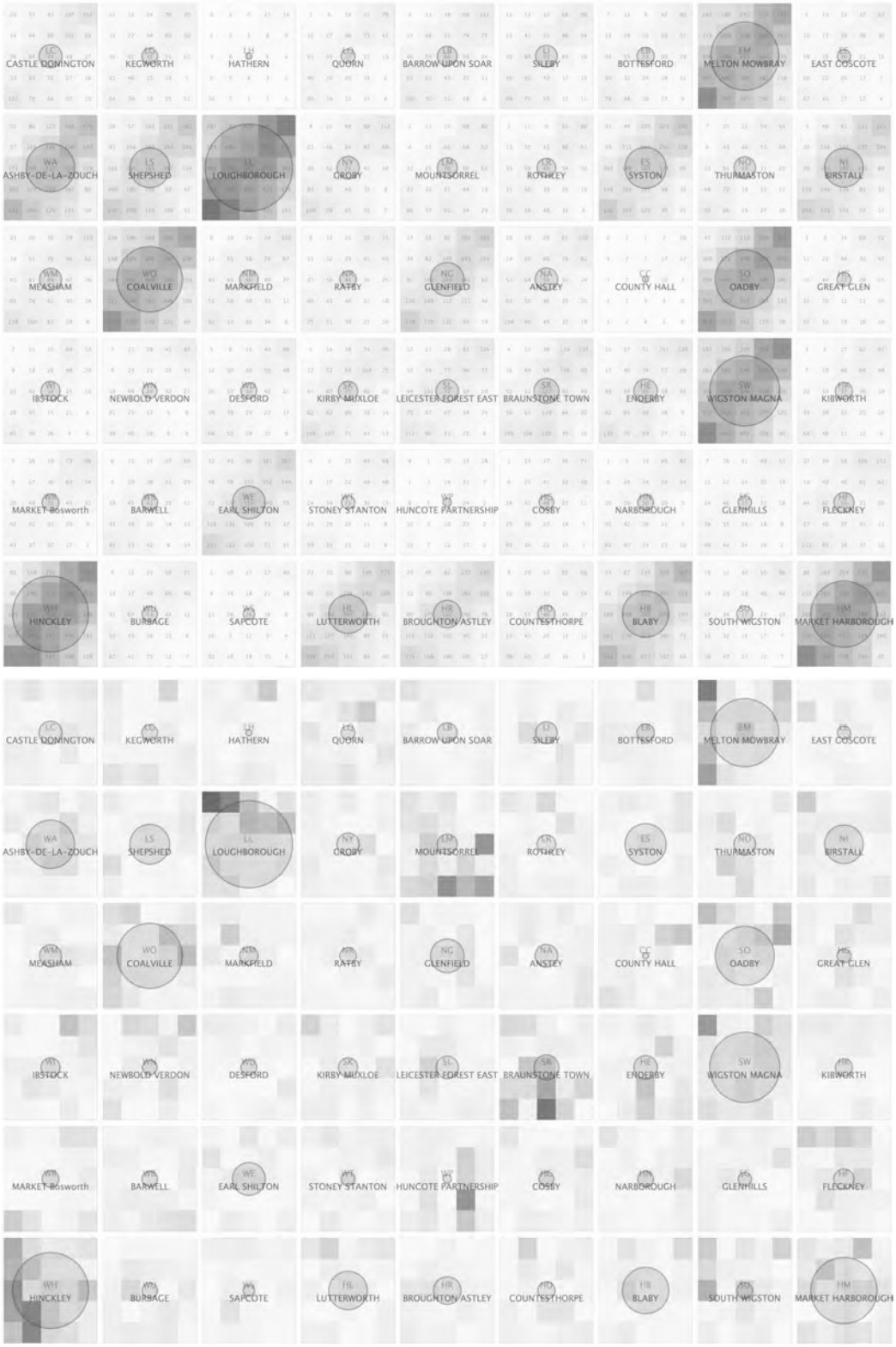



**Figure 5:** RF plots for 54 Leicestershire libraries - fixed size spatial treemap where circle size represents number of registered users per library. Top - numbers of users with scaled global sequential 'YlOrBn' scheme. Bottom - signed-chi statistic with global diverging 'RdBu' scheme (red for higher than expected, blue for lower than expected).

**(iii) How does performance vary across the 54 libraries in Leicestershire?**

**RF / CHI matrix plots:** A matrix showing users in recency (by column) / frequency (by row) quintiles for each library (Figure 5) met LLS needs (see Figure 2, bottom left). The most recent and frequent users are located at the top right of these 'RF plots' (Novos, 2004). Considering RF plots for 54 libraries concurrently was deemed useful by LLS – particularly with animated transitions between spatial and spatially ordered views (Radburn et al., 2009).

Concurrently visualizing signed chi statistics (Figure 5, bottom) reveals high numbers of low recency users in some larger libraries (Melton, Coalville, Loughborough, Wigston) but not others (e.g. Hinckley). Loughborough has more lapsed frequent users than expected in red.

**Stage 6. REFINE – Improve the Representation of the Data**

Feedback on the visualization process was very positive with the LLS Marketing Manager commenting on how "*large complex data can be fully interrogated with different views"*. The speed with which new avenues of analysis could be explored was impressive, vindicating the use of *Processing* as a technology. Functionality developed through iterative refinement included:

**Quartile plots** – concentric distances travelled by the 25%, 50% and 75% closest users (Figure 6) reveal that despite the long-legged spiders most citizens use their local library.

**Standard Ellipse** – summarizing point distributions and indicating the directional pattern of usage. Figure 7 suggests that road and river networks affect spatial usage patterns. Although Birstall and Thurmaston libraries are in close proximity, there is little overlap between best users. This is useful information for LLS when deciding on opening hours as opening one library may not ensure recent / frequent usage by those in the neighbouring catchment.

**Stage 7. INTERACT– Allow Users to Control What They See and How They See It**

Our visualization was an iterative process that involved data users and their expertise continually to drive the analysis. High levels of 'control' were achieved through discussion and rapid development. *Processing* provided considerable scope for developing innovative and interactive graphics quickly and flexibly through access to low level functionality through high-level commands with little hindrance (Dykes, 2005). All three authors were able to develop and share *Processing* code during the visualization process. Interactive software controls were provided at all stages in all prototypes enabling view and focus to be changed through direct manipulation so that the graphical representations introduced here could be accessed (see Figure 1).

## 4. Conclusion

The processes required to manipulate data from acquisition through to visualization are significant. Using *Processing* to apply Fry's visualization model enabled us to iteratively and efficiently ask a huge number of new questions of a large underused data set through interactive graphical means. This has produced some useful hypotheses. But how do we know what to do with the provisional and partial answers that result? We show that visualization has potential in the exploration of local authority data holdings. If visualization of this type, and indeed such large data sets, is to be used effectively to influence and develop policy in organizations that hold them then a focus on an eight stage of visualization may be important: **act**. This is a non-trivial stage with significant social as well as analytical and technical elements that require consideration.

**Acknowledgments**

This work was supported by an ESRC UPTAP User Fellowship RES-163-27-0017 and Leicestershire County Council. Usage data kindly provided by Leicestershire Library Service.

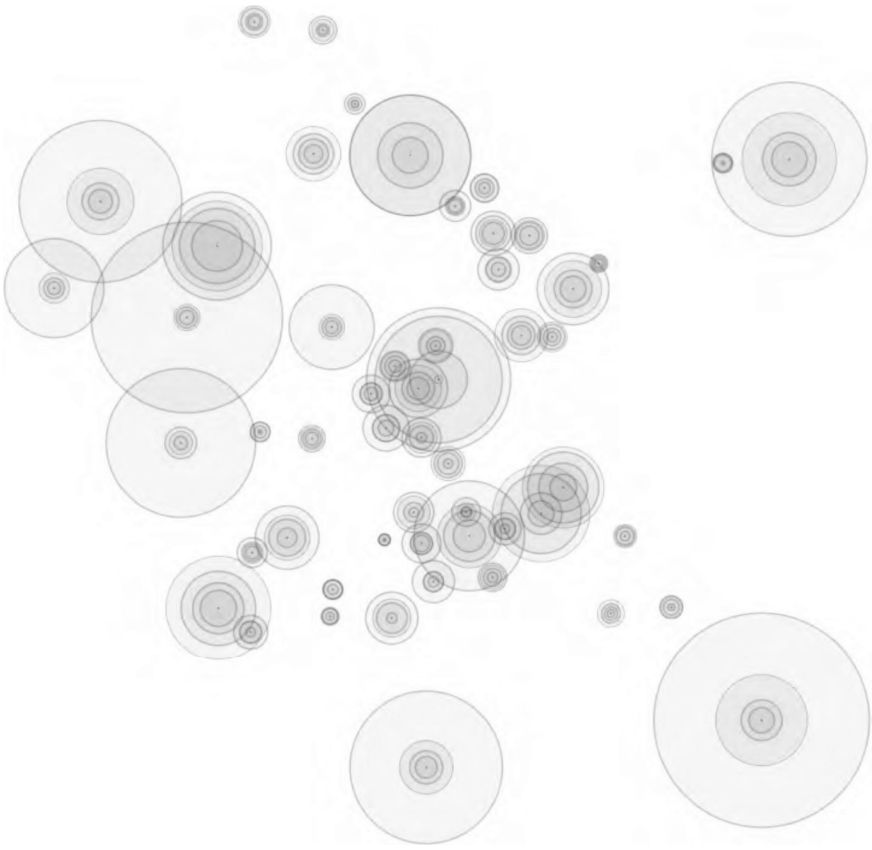


**Figure 6:** 'Quartile plots' for all 54 libraries with geographic locations. Grey symbols are sized by borrower population. Red circles show distances travelled by 25, 50 and 75% of 'best users'.


**Figure 7**: Home locations of 'best users' from neighbouring Birstall and Thurmaston libraries with standard ellipse and 1:50,000 LandRanger.

## References

Dykes, J. (2005). Facilitating Interaction for Geovisualization. In *Exploring Geovisualization*. Amsterdam: Elsevier, pp. 265-291.

Fry, B. (2004). *Computational Information Design*, Massachusetts Institute of Technology.

Fry, B. (2008). *Visualizing Data*. Sebastopol, CA: O'Reilly, 382 pp.

Fry, B., & Reas, C. (2007). *Processing: A Programming Handbook for Visual Designers and Artists*. Cambridge, MA: The MIT Press, 736 pp.

Heer, J. and Bostock. M. (2009). ProtoVis: A Graphical Toolkit for Visualization. *IEEE Transactions on Visualization and Computer Graphics*, **15(6),** 1121-1128.

Meyer, M., Munzner, T. & Pfister, H. (2009). MizBee: A Multiscale Synteny Browser. *IEEE Transactions on Visualization and Computer Graphics*, **15(6)**, 897-904.

Novos, J. (2004). Drilling down: Turning Customer data into profits with a spreadsheet, booklocker.com, 196 pp.

Heer, J., Card, S.K & Landay, J.A. (2005). Prefuse: a toolkit for interactive information visualization, *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, 421-430. ACM New York, NY, USA.

Radburn, R., Thomas, N., Forster, P., and Pye, S. (2007). Beyond the comfort zone. *Public Library Journal*, **22(3)**, 20-23.

Radburn. R., Dykes, J. and Wood. J. (2009). vizLib: Developing Capacity for Exploratory Data Analysis in Local Government – Visualization of Library Customer Behaviour, *Proceedings of GISRUK*, 1-3 April, Durham UK.

Slingsby, A., Dykes, J. & Wood, J. (2009). Configuring Hierarchical Layouts to Address Research Questions. *IEEE Transactions on Visualization and Computer Graphics*, **15(6)**, 977-984.

Vickers, D., Rees, P., and Birkin, M. (2005). *Creating the National Classification of Census Output Areas: Data, Methods and Results*, Working paper 03/3, School of Geography, University of Leeds.

Vickers, D. & Rees, P. (2007). Creating the National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society, Series A*, **170(2)**, 379-404.

Wood, J., and Dykes, J. (2008). Spatially Ordered Treemaps, *IEEE Transactions on Visualization and Computer Graphics*, **14(6)**, 1348-1355.

Wood, J., Dykes, J. & Slingsby, A., (*in press*) Visualization of Origins, Destinations and Flows with OD Maps. *The Cartographic Journal*.

## Biographies

***Robert Radburn*** *was an ESRC UPTAP research fellow at the giCentre, City University London developing capacity for visual exploratory analysis in local government, and is currently Research and Intelligence Team Leader at Leicestershire County Council.*

***Dr. Jason Dykes*** *is a Senior Lecturer at the giCentre, City University London undertaking applied and theoretical research in, around and between information visualization, interactive analytical cartography and human-centred design.*

***Dr. Jo Wood*** *is a Reader in geographic information at the giCentre, City University London with research interests in geovisualization, terrain modelling and object oriented programming for spatial sciences.*

# Exploring the Usability of Geographic Information: A Grounded Theory Analysis

Michael Brown[1], Jenny Harding[1], Sarah Sharples[2]

[1]Ordnance Survey Research, c530, Romsey Road, Southampton, United Kingdom, SO16 4GU
Email: {michael.brown, jenny.harding}@ordnancesurvey.co.uk

[2]Human Factors Research Group, Faculty of Engineering, University of Nottingham,
University Park, Nottingham, NG7 2RD
Email: sarah.sharples@nottingham.ac.uk

KEYWORDS: geographic information, usability, human factors, grounded theory, ergonomics

## 1. Introduction

With the rapid proliferation of Geographic Information Systems (GIS) in recent years GIS users are no longer an all experts; instead they may be a passing user who does not have the time or expertise to learn the intricacies of complex software. With this changing demographic, how we design GIS products must change to acknowledge and allow for all users' needs throughout the design process. This need has been recognised by many in the industry and work has already begun developing user centred GIS interfaces. However, user focused design has yet to influence the design of the Geographic Information (GI) behind the interfaces (Harding et al, 2009).

Hunter, Wachowicz and Bregt (2003) were some of the first to explore this area, identifying the lack of research in this area and presenting a list of elements of GI usability that need to be explored. More recently Harding et al (2009) reviewed some GI case studies that highlight usability issues and propose eleven key research considerations relating to GI usability within three themes: Interfaces; GI Content, Quality, Structure, Formats; and Trust and Value.

This paper describes the analysis of usability issues concerning Ordnance Survey's vector dataset, OS MasterMap®. A Grounded Theory analysis revealed eleven categories and three core themes describing the usability of GI.

## 2. Identifying Usability Problems

A series of studies were performed to identify usability issues with OS MasterMap. Each of these studies collected information from users with a wide variety of GIS experiences, in terms of expertise and types of use.

First, the following Ordnance Survey sources of previously collected data about users were manually searched by a human factors expert for usability issues relating to OS MasterMap.

- A series of 55 interviews with users.
- Six market research reports relating to OS MasterMap.
- Reports produced as the result of two usability workshops with customer support staff.
- A list of technical Frequently Asked Questions (FAQs) collected by customer support staff.
- Three databases containing customer input concerning products.

Next, a heuristic evaluation of OS MasterMap was performed using a cognitive walkthrough technique. This process involved two human factors experts performing a series of common tasks (including data loading, searching, analysis and queries) with the product and evaluating with a list of

pre-defined usability heuristics at key stages of each tasks. Two sets of heuristics were used to analyse the product, one based on research (Gerhart-Powals, 1996) and the other part of an international standard in usability (ISO 9241-110, 2006).

Finally a diary study was carried out with four OS MasterMap users. Each user was asked to complete a diary of any issues encountered relating to the product over a two week period. Once completed the diaries were content analysed in order to discover any usability issues encountered.

A total of 124 unique usability issues were identified by these three methods covering a range of topics from licensing and data delivery to Meta data and data quality. These issues provided a useful tool for the improvement of OS MasterMap, however the large number of issues identified and specific nature of these issues means that they are not easily generalisable to other GI products. For example the issue *"names are repeated across former tile edges, which creates issues of visual appearance when producing output maps for areas crossing former tile edges"* is an artefact specific to OS MasterMap. In order to explore this area further and identify the core usability concepts that underlie these product specific issues a grounded theory analysis was performed on this database of usability issues.

## 3. Grounded Theory Analysis

Based on the guidelines proposed by Corbin and Strauss (2008) and Charmaz (2006), the grounded theory style analysis was performed in four stages. Many types of grounded theory exist, but generally speaking it is a systematic qualitative method focused on the generation of theories based entirely on the data and free of preconceived hypotheses. Whilst it is acknowledged that the usability expertise of the analyst may have affected the nature of code selection, the term grounded theory is retained to emphasise the role of the data in generation of hypotheses, rather than the analysis being conducted using pre-defined codes.

**Table 1.** Sample Results of Grounded Theory

| Issue summary | Code | Category | Theme |
|---|---|---|---|
| The topography layer is not yet used within the system, as a critical need to use it has not yet been identified and also the handling gets a bit complex | Difficulties handling specific layers | Failure to Integrate with Other Data | Lack of User Control over the Data |
| Integrating data can be a frustration (e.g. mapping and census data). It would be good if all spatial data was available as one package which works together | Frustrations integrating multiple data sources | | |
| it would be useful to have some other layers (e.g.) Address Layer 2, Road Routing Information and Points of Interest data as part of the standard package. | | | |
| Need a package for GB data which is reasonable in price (with respect to navigation system pricing) and that includes all data components required (ITN, Addressing, POI – a package that really works with the application). | | | |
| One stop shop for all UK map data (Ireland, Channel islands etc) would be useful | Lack of single data source for British Isles | | |
| It's not easy to understand all the OS licensing terminologies currently used, it's quite difficult to grasp. More clarity and simple terminology about licensing is needed. | Documentation uses overly technical language | Lack of Communication with Users | |
| Much of the OS MasterMap terminology/classifications for features needs to be changed to make sense to non-GIS/technical users | | | |
| Read me file: After the licence information section, the information is very technical and full of jargon | | | |
| OS email sequence before each delivery confusing and off putting: 1. email advising of imminent dispatch. 2. 'order delayed' email. 3. emailed to confirm dispatch | Email order updates are confusing | | |
| The data 'extraction date' is earlier than the 'order date', but no explanation of this is provided | Explanation of updates unclear | | |
| Information in the data on 'reason' for change is not actually meaningful. (Product Managers aware and there are plans to do something about it). | | | |

Firstly the initial data was coded; this process involved describing each issue with a short sentence that conveys the usability concept described by the issue. A single code was produced for most of the issues, but in a few cases it was decided that an issue contained two or more core concepts, so multiple codes were produced. For example, the issue *"vegetation attribution is too general, may be inaccurate, out of date and extents may be out of date"* was deemed to contain three core issues and so it was coded as "*Labels/Classifications not specific enough*", "Data Inconsistency" and "Insufficient Data Updates".

The second stage of the process involved grouping the codes into categories. A method of constant comparison (Corbin and Strauss, 2008) was used to group the codes into categories based on conceptual similarity. These categories were named and each given a single line description. Themes were then created by grouping similar categories together.

The final stage of analysis was to create a theory that describes the data in its entirety.

Table 1 shows a sample of the results of the grounded theory analysis, highlighting the structure of the analysis.

The initial 124 issues produced 64 unique codes. From these codes eleven categories were formed.

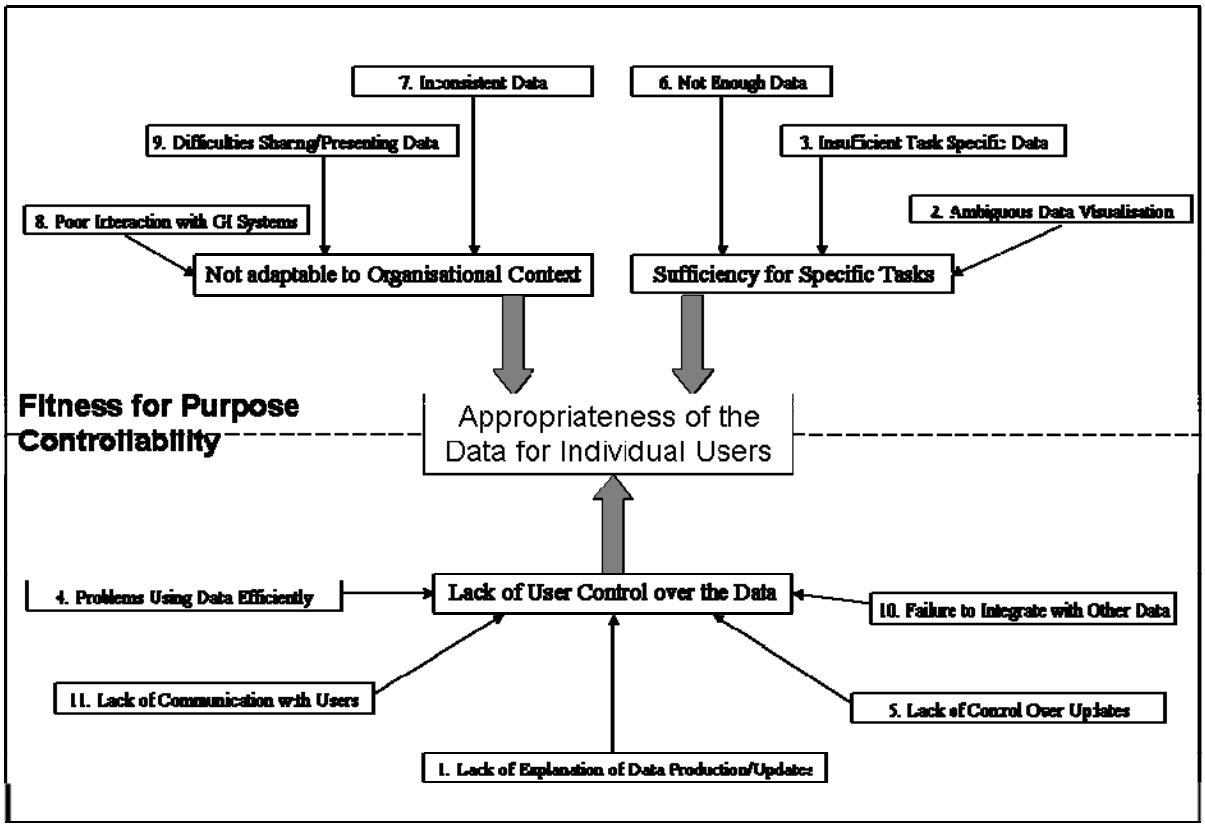
1. *Lack of Explanation of Data Production/Updates:* It is unclear to users why data is placed, stored, labelled or updated in certain ways
2. *Ambiguous Data Visualisation:* The graphic representation of data can be ambiguous
3. *Insufficient Task Specific Data:* Sometimes the detailed information needed to perform specific tasks is not found in the data
4. *Problems Using Data Efficiently:* Working with the data can at times be time consuming
5. *Lack of Control Over Updates:* Users feel like they have little control over the frequency, content and other effects of updates
6. *Not Enough Data:* There can be insufficient labels, attributions or other types of data.
7. *Inconsistent Data:* Some of the data can be inconsistent in terms of labelling, attributions, accuracy or quality.
8. *Poor Interaction with GI Systems:* Users come across a range of problems when trying to use the data with GI systems
9. *Difficulties Sharing/Presenting Data:* At times the data can be difficult to present to, or share with others
10. *Failure to Integrate with Other Data*: The data does not always integrate well with information from other sources.
11. *Lack of Communication with Users:* Communication between the data provider and users is not as clear or frequent as it should be.

From these eleven Categories three themes where formed, each highlighting a key usability issue for GI.

1. *Not adaptable to Organisational Context.* Contains categories 7, 8 and 9.
2. *Lack of User Control over the Data.* Contains categories 1, 4, 5, 10 and 11.
3. *Sufficiency for Specific Tasks.* Contains categories 2, 3, and 6.

The initial theory was identified as "Appropriateness of the Data for Individual Users" and states that "Users want a data product that is appropriate for their individual needs, and allows them to control the data and perform specialist tasks, within a specific organisational context." Table 2 shows how the various categories and themes fed into the construction of this theory. This theory will be further verified and refined through user testing.

**Table 2.** Relationship Diagram of Grounded Theory, Categories and Themes




### 3.6 Conclusions

Two main conclusions can be draw from this analysis. Firstly, the wide range of usability issues discovered highlights the need for human factors methods in the design of Geospatial Information. Secondly, the Categories, Themes and Theory produced show some important areas in which GI usability can be improved.

## 4. Discussion

Comparing our results to the themes proposed by Harding et al (2009) reveals some clear parallels. Each theme is clearly present in at least one of the categories identified in this study, lending more support to the importance of these three key areas. There is also a clear overlap with other work exploring the importance of both fitness for purpose (De Bruin and Bregt, 2001) and controllability (Jameson and Schwarzkopf, 2002) to GIS and usability. Having identified some of the important issues that GI usability must address, the next challenge is developing efficient and effective usability methods that will allow these issues to be identified and fixed during the development of GI products.

## 5. Acknowledgements

The Authors are grateful for the support of Ordnance Survey and the Technology Strategy Board. Also, the contributions of all Ordnance Survey staff involved, especially the work of Emma Pickering is gratefully acknowledged.

## References

Baily, B (1999) Heuristic Evaluations *UI Design Newsletter - May, Human Factors International.* URL: http://www.humanfactors.com/downloads/may99.asp. Accessed Nov 2009.

Charmaz, K (2006) *Constructing Grounded Theory.* Sage Publications Ltd., London

Corbin J and Strauss A (2008) *Basics of Qualitative Research, 3rd Edition.* Sage Publications Ltd., London.

De Briun S and Bregt A (2001) Assessing fitness for use: the expected value for spatial data sets *Geographic Information Science.* 15 (5) pp457-471.

Gerhart-Powels J (1996) Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction* 8(2) pp189-211.

International Standards Organisation (2006). *ISO 9241 - 110: Ergonomics of Human-Computer interaction - Part 110: Dialogue principles.* Geneva ISO

Jameson A and Schwarzkopf E (2002) Pros and Cons of Controllability: An Empirical Study, in *Adaptive Hypermedia and Adaptive Web-Based Systems.* De Bra P, Brusilovsky P and Conejo (Eds). Springer-Verlag, Belin, pp193-202.

Harding J, Sharples S, Haklay M, Burnett G, Dadashi Y, Forrest D, Maguire M, Parker C and Ratcliffe L (2009) Usable geographic information - what does it mean? *Proceedings of the Association of Geographic Information 2009,* Stratford-upon-Avon.

Harvey F and Tulloch D (2006) Local-government data sharing: Evaluating the foundations of spatial data infrastructures *International Journal of Geographical Information Science* 20(7) pp743-768

Hunter J, Wachowicz M and Bregt A (2003) Understanding Spatial Data Usability. *Data Science Journal: Special Data Usability Special Section.* 2 pp79-89.

**Biography**

*Michael Brown is a KTP Associate working with the Ordnance Survey and the University of Nottingham to develop human factors methods for the design of Geographic Information. His primary research interest is in the development of usability design and evaluation methods for application in novel domains.*

*Jenny Harding is a Senior Research Scientist with the Research group of Ordnance Survey. Her main areas of current research include understanding user needs for geographic information and evaluating usability of geographic information products.*

*Sarah Sharples is an Associate Professor in Human Factors at the University of Nottingham. Her main areas of interest and expertise are Human-Computer Interaction, cognitive ergonomics and development of quantitative and qualitative research methodologies for examination of interaction with innovative technologies in complex systems.*

# On Automatic Mapping of Environmental Data Using Adaptive General Regression Neural Network

Mikhail Kanevski, Vadim Timonin

Institute of Geomatics and Analysis of Risk, Amphipôle bldg., University of Lausanne, Switzerland
Tel. +41 22 692 3531
Mikhail.Kanevski@unil.ch, Vadim.Timonin@unil.ch, www.unil.ch/igar

KEYWORDS: environmental data, automatic mapping, artificial neural networks, feature selection

## 1. Introduction

The paper discusses some ideas about the applications of machine learning algorithms for automatic mapping of geospatial data. Mapping is considered as a spatial prediction either of continuous (regression problem) or categorical (classification problem) data. Below mainly the problem of spatial regression/prediction is considered in detail.

An automatic mapping of geospatial data is an important challenge if we consider the number of data available nowadays, the developments of new monitoring networks, and remote sensing data flows (Dubois 2005). In general, geospatial data are not only distributed in a geographical space (two or three dimensional) but should be considered in high dimensional geo-feature spaces, which, for example, can be generated by using digital elevation models. This is a typical case when modelling natural hazards phenomena and environmental risks. The dimension of space can easily achieve ten or more. Some details and real case studies can be found in Kanevski et al (2009) and in Pozdnoukhov et al. (2009). Therefore geostatistics, which is usually based on variogram analysis and modelling, should be replaced by more efficient and appropriate methods. One of the possibilities is based on machine learning algorithms (artificial neural networks, support vector machines, Gaussian processes, etc.).

In this paper adaptive General Regression Neural Networks (GRNN) is considered as a candidate for the tasks of automatic spatial predictions (regression). Correspondingly, Probabilistic Neural Network (PNN) can be considered as a model for automatic spatial classifications (see theoretical details about both models in Kanevski et al (2009). Simplified version of GRNN model considered in this paper was a winner of an international spatial interpolation comparison presented in Dubois (2005).

Let us formulate some general criteria important for automatic modelling. First, it should be able to discriminate between spatial patterns/structures and noise, which is not spatially structured. In the latter case cartography has no meaning. Second, it should be a universal, nonlinear modelling tool - it is desirable that the model is fast, efficient and is automatically tuned/trained. Third, a good model should be able to perform some kind of feature (variable) selection when working in a higher dimensional space.

## 2. Adaptive General Regression Neural Network

General Regression Neural Network is a reformulation of Nadaraya-Watson nonparametric regression model proposed by Specht (1991). Nevertheless, GRNN is more than simple Nadaraya-Watson estimator: in principle, it is not necessary to use data points as centres of the kernels, adaptive models can use different kernels for different inputs, kernels can locally depend on data points, etc.

Let us consider data measurements $Z_n$ ($n=1,\ldots N$) in $m$-dimensional space ($x_1, \ldots, x_m$). According to GRNN model the prediction at unknown point $Z(\boldsymbol{x})$ is defined by the following formula:

$$Z(x_1,...x_m) = \frac{\sum_{n=1}^{N} Z_n \exp\left(-\sum_{i=1}^{m} \frac{D^2(x_i, x_{in})}{2\sigma_i^2}\right)}{\sum_{n=1}^{N} \exp\left(-\sum_{i=1}^{m} \frac{D^2(x_i, x_{in})}{2\sigma_i^2}\right)}$$

where $D^2(x_i, x_{in})$ is a distance between a point of prediction and a measurement $n$, $\sigma_i$ is a kernel bandwidth corresponding to the $i^{th}$ input dimension ($i=1,...m$). In this approximation it is supposed that the interaction between input variables is negligible. In a more general setting full covariance matrix (Mahalanobis distance) can be considered. But already this anisotropic model can take into account many real world phenomena and makes the discrimination between important and very noisy or not relevant inputs (see below).

The only unknown parameters in GRNN model are kernel bandwidths, and this is a topic of GRNN training. These parameters can be tuned by splitting data into training and validation subsets and then by minimizing the validation error with correspondingly chosen bandwidths or (if there are not too many data) by a simple cross-validation or leave-one-out criterion. Gradient search from a good starting point ("warm start"), for example estimated by an isotropic kernel model, is another fast and efficient solution to minimize the cross-validation error in anisotropic case.

Let us remind some properties of GRNN which are useful for automatic mapping: 1) for isotopic kernel and cross-validation training the solution is unique; 2) when there are no spatial structures, i.e. only spatially not correlated noise, there is no minimum on cross-validation curve and kernel bandwidth tends to infinity (it is larger than the region of the study, i.e. maximal distances between data points). This is also true when only some coordinates (input variables) are very noisy and useless for the prediction. For example, when $k^{th}$ -input variable is a noise the value of the corresponding bandwidth $\sigma_k \to \infty$. In this case

$$\forall n: \quad \frac{D^2(x_k, x_{kn})}{2\sigma_k^2} \ll 1, \quad \text{and, correspondignly,} \quad \exp\left(-\frac{D^2(x_k, x_{kn})}{2\sigma_k^2}\right) \approx 1$$

which means that input $k$ makes no contribution. Therefore, adaptive GRNN performs automatic feature selection, i.e. non relevant inputs are automatically neglected.

This phenomenon was also observed in Specht and Romsdahl (1994) and in a weighted k-NN feature selection algorithm in Navot et al (2005), which actually corresponds to the anisotropic GRNN model. More details on theory, training and applications of GRNN for spatial data, can be found in Kanevski et al (2009).

Moreover, GRNN model can be used to analyse the residuals of spatial data modelling using different algorithms, like multilayer perceptrons, support vector machines etc. in order to control the quality of their results: the residuals of optimal models should be white noise without any spatial structures. And this completes the automatic modelling procedure: from raw data analysis via modelling to the analysis of the results and the residuals.

### 3. Case studies

Let us demonstrate the ideas presented above using simulated and real case studies. The precipitation data in Switzerland are used to check the hypotheses presented above. Both original data and shuffled data (randomized) with destroyed spatial structure were applied. The demonstrative results are shown in Figure 1. Upper figures correspond to original data of precipitation modelled in a 3-dimensional space (X = longitude, Y = latitude, and Z = altitude). With isotropic model, a well-defined minimum on cross-validation curve is observed, while for shuffled precipitation data (bottom) there is no minimum – no spatial structure.

For preparing the final map on a dense grid an anisotropic model was trained (see basic parameters in Table 1) and the result for the correct inputs is given in Figure 2 (left). The result of Model2 – 4D model with additional wrong coordinate (in fact shuffled altitude was used for the interpretability reasons) are given in Table 1 and a corresponding map in Figure 2 (right). Let us remind that the difference between min and max values in altitude in Swiss Alps is about 4100 m. Both results are very similar which corresponds to theoretical considerations presented above. In fact, it means that GRNN automatically discarded the wrong input variable.

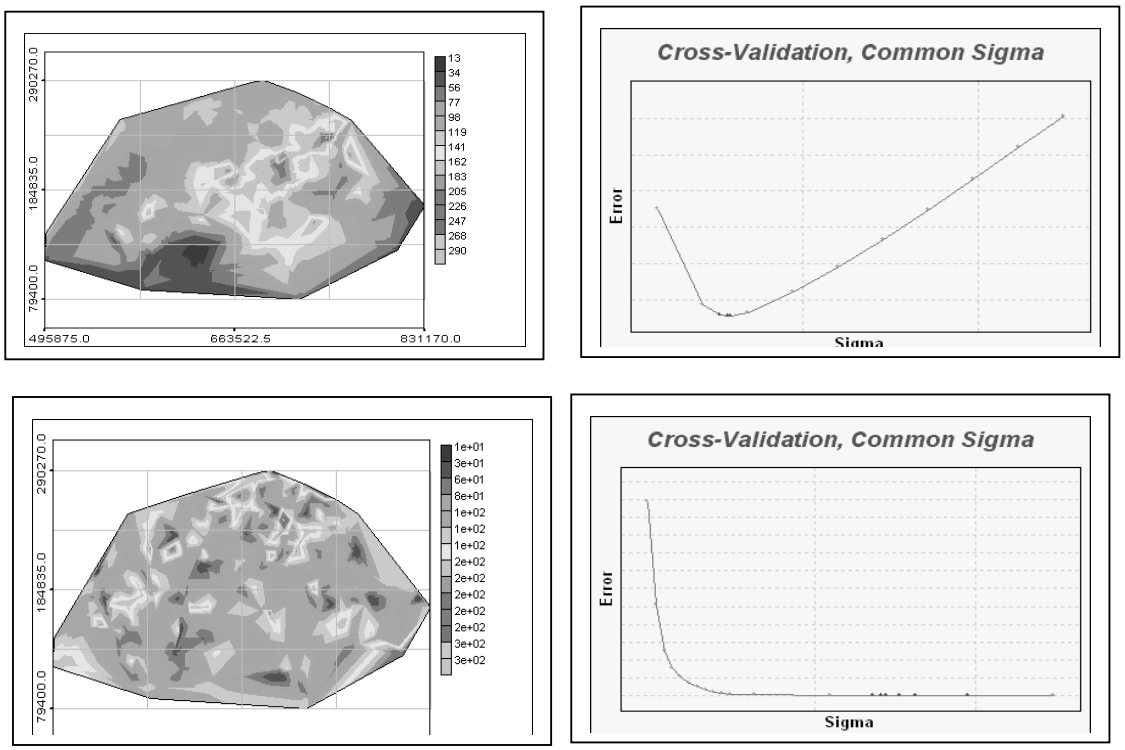



**Figure 1.** Patterns of precipitation (left) and corresponding isotropic cross-validation curves (right): original data (up), shuffled data (bottom). Linear interpolation of data is used just for the visualisation purposes (left maps of isolines were prepared using Delaunay triangulation).

**Table 1.** Comparison of kernel bandwidths for 3D and 4D models.

| Model | Cross-Validation error | Sigma values (metres) | | | |
|---|---|---|---|---|---|
| | | $\sigma_x$ | $\sigma_y$ | $\sigma_z$ | $\sigma_{Znoise}$ |
| 3D | 419 | 7011 | 7601 | 192 | |
| 4D (3D+Noise) | 420 | 6949 | 7474 | 191 | 4135 |

An interesting topic deals with a software (operational) implementation of this approach. When the number of input variables d is not very high (let us say less than 15) and the number of data is about hundreds-thousands all combinations of the variables (= $2^d$) can be considered and corresponding cross-validation errors can be estimated in order to find the best solution by using an adaptive GRNN. When the number of inputs and data is larger a simulated annealing or genetic algorithms can be applied to find acceptable solutions in a reasonable time.

## 4. Conclusions

GRNN is proposed as an efficient tool for automatic exploratory spatial data analysis (raw data and residuals) and spatial predictions. Discriminative properties of GRNN were demonstrated using real and simulated data. A test on input variables selection by GRNN was performed and promising results were obtained. The future research will be concentrated on studying GRNN in higher dimensional geo-feature spaces and as a feature selection tool. Scaling of the algorithms according to the dimension of the input space and new challenging case studies using environmental and natural hazards data are interesting topics of the current research. The main results were prepared using Geostat Office (Kanevski and Maignan, 2004) and a GRNN model implemented in a Machine Learning Office (Kanevski, Pozdnoukhov and Timonin, 2009) available for scientific community.

## 5. Acknowledgements

The research was partly supported by Swiss NSF grants N 200021-126505 and N 200020-121835.

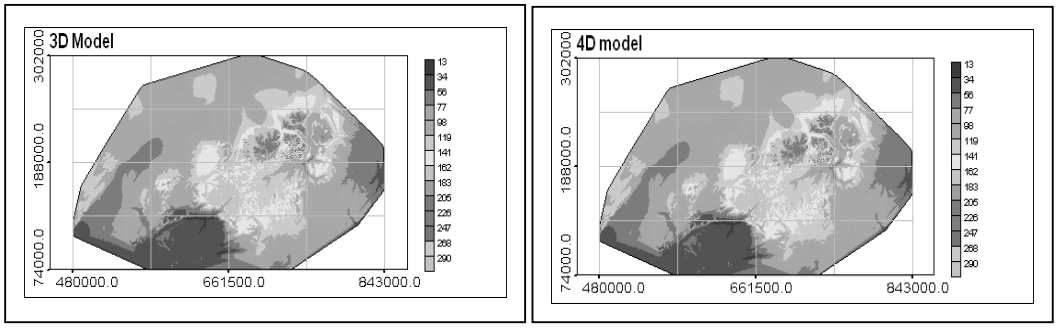



**Figure 2.** GRNN mapping of precipitation using 3D and 4D (=3D+noise) models.

## References

Dubois G. (Editor) (2005) *Automatic Mapping Algorithms for Routine and Emergency Data*. EU Report No. EUR 21595 EN.

Kanevski M., Pozdnoukhov A. and Timonin V. (2009) *Machine learning for spatial environmental data. Theory, applications and software.* EPFL and CRC Press.

Kanevski M. and Maignan M (2004) *Analysis and Modelling of Spatial Environmental Data.* EPFL and Marcel Dekker Press.

Navot A., Shpigelman L., Tishby N., and Vaadia E. (2005). Nearest Neighbour Based Feature Selection for Regression and its Application to Neural Activity. *NIPS2005 conference*.

Pozdnoukhov A., Foresti L. and Kanevski M (2009) Data-driven topo-climatic mapping with machine learning methods. *Natural Hazards*, 50 pp. 497-518.

Specht D. (1991) A General Regression Neural Network. *IEEE Transaction on Neural Networks*., 2 pp. 568-576.

Specht D. and Romsdahl H. (1994) Experience with Adaptive Probabilistic Neural Networks and Adaptive General Regression Neural Networks. *Proceedings of the IEEE World Congress on Computational Intelligence,* 2 pp. 1203-1208.

**Biography**

*Prof. Mikhail Kanevski is a chair of geomatics at the Institute of Geomatics and Analysis of Risk, University of Lausanne. His scientific interests include: geospatial data and time series analysis and modelling, geo-visualisation, remote sensing and satellite images classification, geostatistics, application of machine learning algorithms for geo- and environmental data, natural hazards and risks, analysis of socio-economic and financial data.*

*Vadim Timonin is a PhD student at the Institute of Geomatics and Analysis of Risk. His research topics cover development of models and algorithms for automatic mapping of environmental data using machine learning algorithms. He is a main contributor to Machine Learning Office.*

# Automatic Classification of Retail Spaces from a Large Scale Topographic Database

William A Mackaness[1], Omair Chaudhry[2]

[1]School of GeoSciences, University of Edinburgh, Drummond Street, Edinburgh, EH8 9XP
Tel: 0131 650 8163 | Email: william.mackaness@ed.ac.uk

[2]Manchester Metropolitan University John Dalton, Chester Street, Manchester, M1 5GD
Email: O.Chaudhry@mmu.ac.uk

KEYWORDS: retail, classification methodologies, fuzzy logic, Bayesian,

## 1. A need to classify retail spaces

There is considerable interest in understanding distribution patterns of different types of retail space at a national scale. Retail classification is 'essential as a means of understanding and analysing relationships in the work of retailing' (Guy 1998, p255). It allows us to observe changing patterns of distribution and an understanding of consumer behaviour. Many factors govern urban morphology (Schillers 2001); many measures can be used to define different types of retail space (Guy 1998). How meaningful these metrics are depends to some degree on the reason for classification, be it asset management, measure of service provision, government policy or for retailers themselves (Pitt and Musa 2009). In this paper we argue that form follows function, and that many of the characteristics of different retail spaces are manifest in the extent, and patterns of distribution of retail buildings. When coupled with additional information (such as access and transport information, parking areas and retail type), this paper illustrates that it is possible to automatically and systematically classify retail spaces using fine scale topographic data. We begin with a discussion of the characteristics of various retail spaces; we describe how various measures can be modelled using a variety of national coverage datasets. We contrast three methodologies that use these metrics in various ways to generate various classifications. The outputs of the analysis are compared in order to assess the efficacy of these approaches.

## 2. Characterising Retail Spaces

The city is comprised of different types of retail space, varying in density, composition, and location – from 'the high street' and the shopping mall to the retail park and the factory outlet (Figure 1).

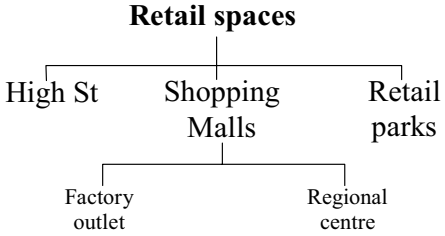



**Figure 1:** A hierarchy of types of retail spaces found in urban areas

### *2.1 'The High Street'*

Typically the High street contains a multiplicity of owners, with a predominance of retail outlets, clustered in an unplanned manner, along arterial roads – originally 'seeded' from periodic markets that once served the local community. Benefits include ease of access via public transport, and clustering of shops along the high street, which may additionally provide parking.

### *2.2 The Shopping Mall*

The Shopping Mall is defined as 'many units of shops but managed as a single property' (Pitt and Musa 2008), and comes in many guises (DeLisle 2007) – a Regional Shopping Centre being a large version of the Shopping Mall. The Shopping Mall may form an adjunct to the High Street, and is typically well served by public transport. Thus the key defining characteristics from a morphological point of view, is that they often lie within a city block, and have a high density of shops contained within one building.

### *2.3 Retail Parks*

Retail Parks are 'loose groupings of superstores and retail warehouses located at nodal points in the suburbs' (Bromley and Thomas 1988, p4). They are often adjacent to major arterial roads or junctions with access focused on the car. Factory Outlets follow a similar profile. Retail parks are found at the edge of town, where lower rental values result in expansive, single storey outlets. Table 1, summarises key characteristics identified from the literature.

Table 1: Defining characteristics of a subset of retail spaces (variously sourced from Guy 1998; and Schillers 2001).

| **Retail type** | **Access** | **Description & Function** | **Gross area** | **No. of stores** | **Centrality** |
|---|---|---|---|---|---|
| High Street | Pedestrian and Bus, may include on street parking, localised car parks | Unplanned, loose clustering along 'important streets' (shopping ribbons), possibly street parking. | | | Town centre, out-of-centre in larger cities |
| Shopping Mall | Pedestrian and bus. May include underground car park. | Single building, planned, with Anchor store, plus smaller retailers. May include leisure uses. Typically within a city block. Often found in the High Street. | >10,000 $m^2$ and < 37,000 $m^2$ | >=50 retail stores. | Often city centre, co-located with 'high street'/ central shopping areas. |
| Regional Shopping Centre | Multiple road access, few buses. Shared parking area, pedestrian paths. | 'large'shopping mall (single building, planned). May be over two floors. Typically within a city block. | Can range from 37k $m^2$ – 74k $m^2$ | >50 retail stores | 'off centre' outside town centre |
| Factory Outlet Centre | Pedestrian and bus service. | 'small' shopping mall (clustered, planned). Typically within a street block. No anchor stores. Discounted goods. | >10k $m^2$ . | >50 retail stores. | Out-of-centre, edge of centre. |
| Retail Park | Multiple road access, few buses. Shared parking area, pedestrian paths. | Cluster of several large stores, including anchor stores, planned, incl. selling bulky items. Typically within a city block. May include leisure uses (cinema, indoor bowling). | >15240 $m^2$ gross lettabe area. | > 3 warehouses, each >3048 $m^2$ at least. | Out-of-centre, edge-of-centre |

## 3. Measures to discern different Retail Spaces

### *3.1 Form, Composition and Extent*

The information used to measure the various characteristics was sourced from various 'layers' within Ordnance Survey's MasterMap product. The Address layer provided information on the type of shop and the number of shops within a single building.

### *3.2 Urban Centrality*

Among a number of defining characteristic of retail spaces is their location relative to the centre of town. The centrality measure, at any given location 'x', was devised by fitting a convex hull to the urban extent (Chaudhry and Mackaness 2008) and normalising the value between the edge of the urban extent and its centre (OS Meridian 2)(Figure 2).

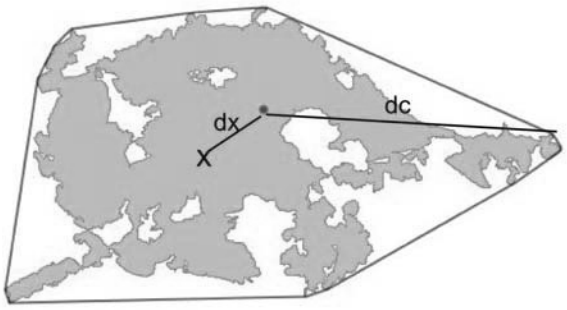



**Figure 1:** A measure of centrality

### *3.3 Accessibility*

Retail spaces are served by a mix of public and private transport. Central sites benefit from the hub effect of bus services, whilst edge of centre sites suffer poorer bus services, instead catering more for the car, with large shared parking. The density of bus stops at the location of retail spaces was determined using PointX data (Ordnance Survey 2009b). The density of roads servicing a particular site can be determined from road network data (Ordnance Survey 2009a). The density for both datasets is determined by using kernel density estimation (Wasserman 2005) (Figure 3).

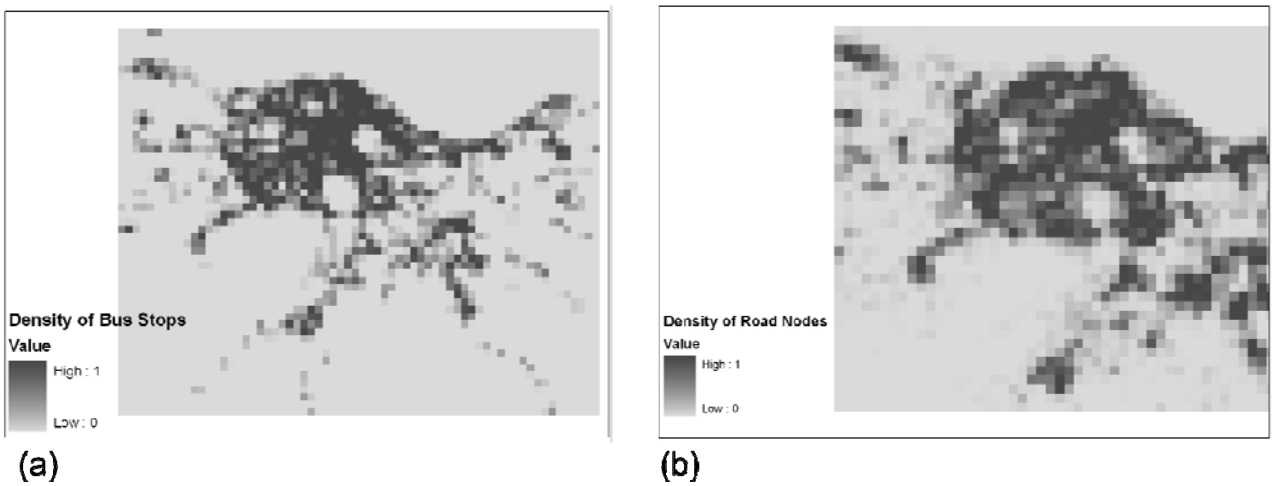



**Figure 2:** (a) Density of bus stops in Edinburgh city (b) density of OS ITN road nodes in Edinburgh city

By using these various metrics, we were able to model five of Guy's (1998) ten suggested dimensions (namely function, size of store, physical form, location and development type).

## 4. Identifying Retail Spaces

But before we can begin the process of classification, we must first identify the retail spaces themselves. Determining the extent of a site was based on previous work (Chaudhry et al. 2009). Reflecting on the adage that 'function defines form', the algorithm is able to group together different features according to their shared function. Figure 4 illustrates how the algorithm has grouped together various features (a), and identified an aggregated retail space (b).


**Figure 4:** (a) Selected features that belong to a retail space (b) Aggregated geometry of the retail space from features in (a). (c) Representation of this retail space at 1:25,000 scale. (Ordnance Survey © Crown Copyright. All rights reserved)

## 5. Three Alternate Methodologies – Boolean, Fuzzy and Bayesian

The next challenge was to classify these spaces (Figure 4b) according to a hierarchy (Figure 1) using various metrics (section 3.0). The detection and classification of 'high streets' was based upon work by Chaudhry et al (in press). The approach involved identifying clusters of commercial buildings using minimum spanning trees (Cormen et al. 2001; Regnauld 2003) and combining them with roads (Thom 2005; Thomson and Brooks 2007; Chaudhry and Mackaness 2005). By identifying continuous lengths of road, associated with clusters of retail outlets, it was possible to identify 'the high streets' of a city (Figure 5).

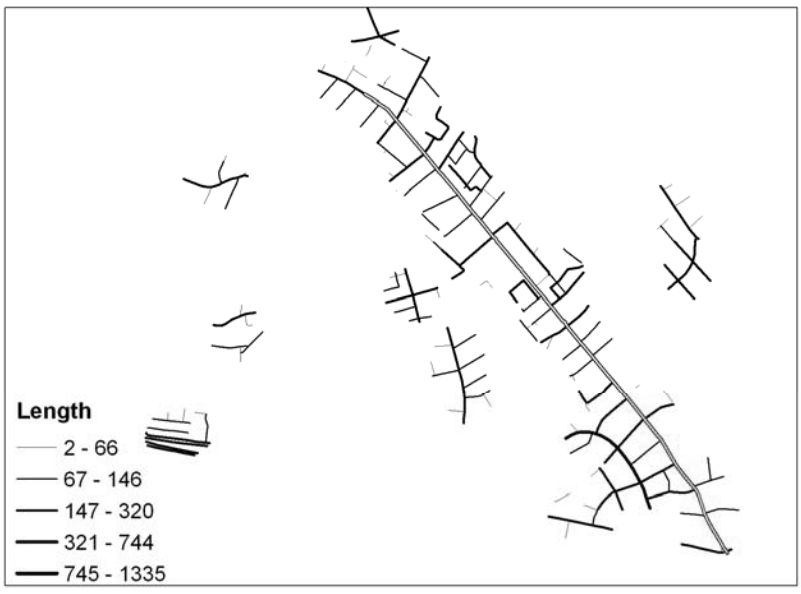



**Figure 5:** The highlighted road (about 1000m long) is that of Shirley high street in Southampton, UK.

For the remaining retail spaces one can envisage a number of different methodologies for classifying retail spaces, using various metrics. A simple approach would be to determine, for each site, whether a particular characteristics was present or not (section 5.1). Alternatively a range of values for each metric could be considered, affording more flexible definitions of different retail spaces (section 5.2). A third approach is to take a set of retail spaces for which their classification is known, and to compare an unknown retail space against the known sample and measure how similar or dissimilar it is, as a way of classifying that space (section 5.3). Each approach was applied to – Edinburgh, Glasgow and Southampton (cities in Great Britain).

*5.1 Boolean Logic*

In the first approach we used crisp definitions and thresholds for defining shopping malls and retail parks. In this type of inference, a retail space is a shopping mall or not (0 or 1).

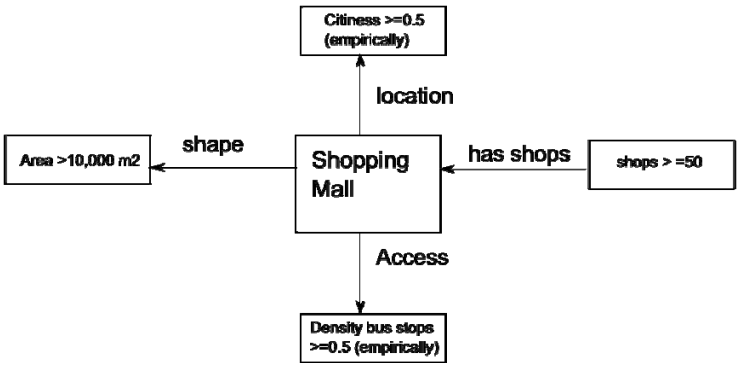



**Figure 6:** Criteria used to define a shopping mall.

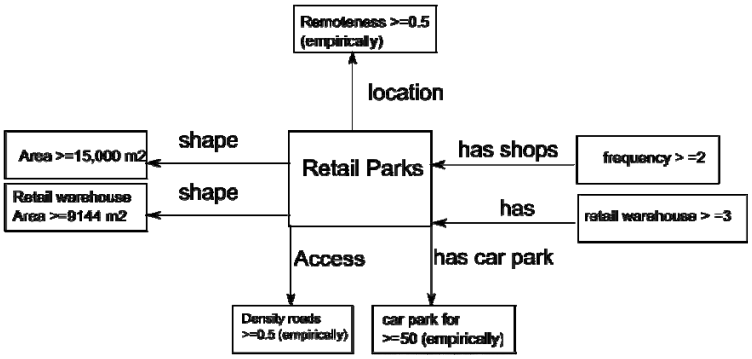



**Figure 7**: Criteria used to define a retail park.

Table 2 shows the results for the three cities, using these crisp definitions of a shopping mall and a retail park. The candidate sites in tables 2, 3 and 4, refer to retail spaces generated by the approach outlined in section 4. Each retail space has at least one building classified as a shop by Address Layer. Table 2 compares the automatically classified results against manually identified shopping malls and retails parks for the three test areas. The results are very poor, perhaps reflecting Guy's observation that the classification of retail spaces lie along a continuum, and that each type of space can vary in it character, and may not have all the attributes that we typically associate with a particular retail space.

**Table 2:** Classification based on a Boolean approach to classification of retail spaces.

| | #Candidate Sites | #Manually identified Shopping Malls | #Correctly identified | #incorrectly identified | #Correctly identified without empirical set thresholds | #incorrectly identified without empirical set thresholds |
|---|---|---|---|---|---|---|
| Edinburgh Shopping Malls | 17 | 6 | 0 | 0 | 1 | 0 |
| Edinburgh Retail Parks | 19 | 6 | 0 | 0 | 0 | 0 |
| Glasgow Shopping Malls | 36 | 6 | 0 | 0 | 0 | 0 |
| Glasgow Retail Parks | 28 | 6 | 0 | 0 | 0 | 0 |
| Southampton Shopping Malls | 6 | 2 | 0 | 0 | 0 | 0 |
| Southampton Retail Parks | 8 | 4 | 0 | 0 | 0 | 0 |

### *5.1 Fuzzy Logic*

Using fuzzy logic acknowledges the vaguee nature of some of these characteristics. It is an approach that explicitly acknowledges that we have prototypical views of these different retail spaces.Iinstead of using crisp thresholds (0 or 1), we use normalised values (0 -1) for each measure. For instance instead of assigning 0 to shop count for a retail space with 25 shops (less than the threshold of 50) we assign it a value of 0.5 – dividing the actual value (25) by the threshold (50) (eg Figure 8). We can do this normalisation using the actual values for each measure together with the threshold values (Figure 6 and 7).

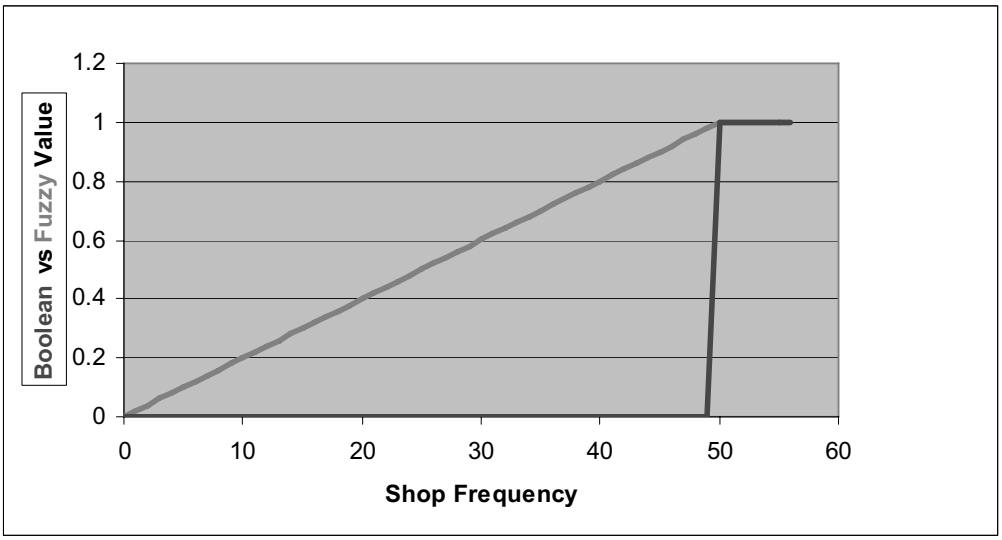



**Figure 8:** Fuzzy and boolean values against the shop frequency.

In order to classify a retail space into a shopping mall or a retail park, we need to combine the values from different measures into an overall value. Weighted linear average (as proposed by Luscher et al. 2008) was used. We calculate the degree of congruence between reality and the ideal prototype using equation 1.

$$con(C_{i,}R_k) = (\sum w_j con(C_{j,}R_k)) / \sum w_j \qquad \text{Eq. 1}$$

Where *con(Cj, Rk)* is the congruence value of a constituent concept of *Ci* and the weight *wj* is an influence value of the subconcept. Initially all weights were equated to 1. *con(Ci, Rk)* = 0 when a realisation *Rk* differs completely from a template *Ci*, and *con(Ci, Rj)* = 1 when they match perfectly. This approach correctly identifies many of the shopping malls and retails parks in the three test areas (Table 3). But there are quite a few omission and commission errors. Commission errors are especially significant in the case of shopping malls. This is due the fact that all the factors are given equal weights thus if there is building with just 1 shop but with high centrality and accessibility, it will be classified as a shopping mall. By re-running the algorithm, the weights (in effect, the importance attached to each metric) were adjusted in order to improve the quality of classification – seeking to minimise errors of commission and omission. The results show a close correlation with a manual classification (Table 3 last two columns). Weights were set experimentally.

**Table 3:** Classification using a fuzzy logic approach.

| | #Sites | #Manually identified | #Correctly automatically identified with all wts =1 | #incorrectly automatically identified with all wts =1 | #Correctly automatically identified with wts | #incorrectly automatically identified with wts |
|---|---|---|---|---|---|---|
| Edinburgh Shopping Malls | 17 | 6 | 5 | 8 | 5 | 1 |
| Edinburgh Retail Parks | 19 | 6 | 3 | 0 | - | - |
| Glasgow Shopping Malls | 36 | 6 | 6 | 8 | 3 | 0 |
| Glasgow Retail Parks | 28 | 6 | 5 | 1 | - | - |
| Southampton Shopping Malls | 6 | 2 | 2 | 0 | 2 | 0 |
| Southampton Retail Parks | 8 | 4 | 1 | 1 | - | - |

### *5.3 Bayesian Inference*

A third approach was to use Bayesian inference which obviates the need to normalise and weight the metrics. Bayes' Rule is a simple way of calculating conditional probabilities (Hacking 2001). Using a Bayesian approach, we can answer questions of the following form: 'For a given, unclassified retail space with a specific set of characteristics, what is the likelihood that it belongs to the population of 'shopping malls' with their specific set of characteristics?'. What is returned is a probability value reflecting the likelihood that the unclassified site is indeed a shopping mall. We used an approach similar to that proposed by Leusher et al. (2009). The joint conditional probability for a classification of an unknown is given by equation 2.

$$P_c(\vec{f} \mid C = c) = \frac{1}{N \mid\mid \vec{h} \mid\mid} \sum_{i=1}^{N} K\left(\frac{\vec{f} - \vec{f_i}}{\vec{h}}\right) \qquad \text{Eq 2}$$

Where Pc is the conditional probability of the unkown for the predicted class C, N is the

number of samples in the training dataset, $\vec{h}$ are the bandwidths, K is the standard normal distribution function, $\vec{f}$ is the vector of properties (Figure 6, 7) of unknown, $\vec{f_i}$ is the vector of the same properties of training dataset.

Here in this research the manually classified shopping malls and retails parks for Edinburgh city were used as our training dataset . Once trained it was used to classify shopping malls and retail parks in Glasgow and Southampton. The results shown in Table 4 show that there are very few cases of omission and commission errors as compared to previous to approaches.

**Table 4:** Classification results using a Bayesian approach.

| | #Candidate Sites | #Manually identified | #Correctly automatically identified | #incorrectly automatically identified |
|---|---|---|---|---|
| Glasgow Shopping Malls | 36 | 6 | 4 | 1 |
| Glasgow Retail Parks | 28 | 6 | 6 | 1 |
| Southampton Shopping Malls | 6 | 2 | 2 | 0 |
| Southampton Retail Parks | 8 | 4 | 3 | 1 |

## 6. Conclusion

DeLisle (2005, p2) talks of a 'dynamic tension between too few and too many classes' in retail space classification arguing that the basis of a sound system of classification is one that has metrics that are unambiguous, meaningful and measurable. Where definitions can be agreed, it is possible to automate the process of identifying and classifying different retail spaces. This paper has proposed a frame of reference that could be used as a basis for painting a national picture, as well as help guide any re-assessment of those criteria. Each methodology has its strengths and limitations in terms of clarity, ease of use, and data requirements. These techniques can produce a systematic classification of retail spaces, through the uniform application of meaningful criteria. The evaluation indicates the correctness of the approach in seeking to minimise classification error. The work also illustrates the breadth of utility afforded through the use of Ordnance Survey data – in particular MasterMap data, and how such databases can be enriched through the use of various analysis techniques.

## 7. Acknowledgement

Authors are grateful for funding from Knowledge Transfer Partnership 6837.

## 8. References

Bromley, R.D.F and Thomas, C.J. 1988 'Spatial and Functional Integration of Retail Units in the Swansea Enterprise Zone', Transactions of the Institute of British Geographers, 13(1): 4-18.

Cormen, T.H. Leiserson, C.E. Rivest R.L. and Stein, C. 2001. 'Chapter 23: Minimum Spanning Trees', Introduction to Algorithms. MIT Press and McGraw Hill. Pp 561-579.

Chaudhry, O. Z. and Mackaness, W.A. (2005) Rural and Urban Road Network Generalization Deriving 1:250,000 From 1:1250. In International Cartographic Conference pp. 10p. CD-ROM: Theme 9: Cartographic Generalisation and Multiple Representation, Session 4, La Coruña- Spain. 10-16 July.

Chaudhry, O. Z. and Mackaness, W.A. (2008) Automatic Identification of Urban Settlement Boundaries for Multiple Representation Databases. Computer Environment and Urban Systems, 32(2), pp. 95-109.

Chaudhry, O. Z, Mackaness, W.A. and Regnauld, N., (2009) A Functional perspective on Map Generalisation, Computer Environment and Urban Systems: Special Issue on Geo-Information

Generalisation and Multiple Representation, 33 (5), pp. 349-362

DeLisle, J.R. 2005 Shopping Centre Classification: Challenges and Opportunities. ICSC Working Paper Series. www.**icsc**.org/srch/rsrch/wp/USSC_Class_091305.pdf

DeLisle, J.R. 2007 ICSC Working Paper Series: Shopping Centre Classifications. New York: International Council of Shopping Centres.

Gravelle, M. Chaudhry, O.Z. and Regnauld, N. 2009 Automatic detection and classification of building clusters based on their function. (Submission to GISRUK2010)

Guy, C.M. 1998 'Classifications of retail sotres and shopping centres: some methodological issues' GeoJournal 45: 255- 264.

Guy C.M. 1994 The Retail Development Process Routledge London.

Hacking, I. (2001). *An introduction to probability and deductive logic.* Cambridge, England: Cambridge University Press.

Lüscher P., Weibel, R. and Mackaness, W.A. 2008 Where is the Terraced House? On the Use of Ontologies for Recognition of Urban Concepts in Cartographic Databases. In Headway in Spatial Data Handling. Lecture Notes in Geoinformation and Cartography. Berlin Heidelberg: Springer , pp. 449-466

Lüscher P., Weibel, R and Burghardt, D. (2009) Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data, Computer Environment and Urban Systems: Special Issue on Geo-Information Generalisation and Multiple Representation, 33 (5), pp. 363-374

Ordnance Survey 2009a Integrated Transportation Network Data http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/itn/ accessed 20.11.09

Ordnance Survey 2009b PointX Data http://www.ordnancesurvey.co.uk/oswebsite/products/pointsofinterest/ accessed 20.11.09

Ordnance Survey 2009c Address Layer 2 http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/layers/addresslayer2/ accessed 20.11.09.

Pitt, M. and Musa, Z.N. 2009 'Towards defining shopping centres and their management systems' Journal of Retail and Leisure Property 8: 39-55.

Regnauld, N. (2003). Algorithms for the amalgamation of topographic data. In: Proceedings of the 21st International Cartographic Conference, Durban, South Africa

Schillers, R 2001 The Dynamics of Property Location. Spon Press, NY.

Thom, S. (2005), A Strategy for collapsing OS Integrated Transport Network (ITN) dual carriageways, presented at the 8th ICA Workshop on Generalisation and Multiple Representation, La Coruña, Spain

Thomson, R. C., and Brooks, R. (2007), Generalisation of Geographic Networks, In Generalisation of Geographic Information: Cartographic Modelling and Applications (eds W.Mackaness, A. Ruas & L.T. Sarjakoski), pp. 255-268. Elsevier, Oxford

Wasserman, L. (2005). *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics.

**Biography**

*William Mackaness is a senior lecturer in the School of GeoSciences, at The University of Edinburgh. His research interests lie in the modeling of geographic phenomenon at multiple scales, and visualisation methodologies*

.

*Omair Chaudhry is concluding his Associate role in a Knowledge Transfer Project jointly funded by Ordnance Survey and is about to take up a lectureship at Manchester Metropolitan University. His interests are in multi scale semantic modelling.*

# Trust in Web GIS: A Preliminary Investigation of the Environment Agency's WIYBY website with non-expert users

Artemis Skarlatidou[1], Muki Haklay, Tao Cheng, Nicola Francis

[1]Department of Civil, Environmental and Geomatic Engineering, University College London
Gower Street, London, WC1B 6BT
Tel. +44 20 7679 2741 | Email: a.skarlatidou@ucl.ac.uk

KEYWORDS: Human-Computer Interaction, usability, aesthetics, trust, user experience, Web GIS

## 1. Introduction

The World Wide Web (Web) provided the basis for wider public access to spatial information and knowledge. Since the introduction of Xerox PARC Map Viewer, there is a high growth in the number of Web GIS applications for public use in different contexts. Several people rely on these systems to provide them with instructions (e.g. find a place) or, in other cases, with more sophisticated tools to perform spatial analysis. An example of such an application is the "What's In Your Back Yard" (WIYBY) website provided by the Environment Agency (EA), which allows users to explore environmental data using a Web GIS interface.

The democratisation of GIS after the establishment of the Web had as a result these systems to be used by people without any GIS experience or knowledge (Haklay and Zafiri, 2008). These non-expert users have a limited understanding of spatial data handling, which introduces uncertainty that is further increased due to the complexity of Web GIS interfaces (Skarlatidou and Haklay, 2006). Moreover, uncertainty and complexity are inherent in the context that these applications are used, for example Web GIS used to investigate site selection problems. Although, uncertainty, risk and dependence are trust preconditions and at the same existent in Web GIS, no one yet considered to investigate trust in this context.

Online trust is well-researched from a Human Computer Interaction (HCI) perspective, especially for the e-commerce domain. Existing studies suggest that people's trust perceptions about electronic online environments, influence the intentions to engage, the perceived User Experience (UXP), the use and acceptance of these systems, and enhance cooperative behaviours (e.g. Schneiderman, 2000). These elements are important for Web GIS, and thus it is critical to investigate how the trustworthiness can be improved.

Existing studies suggest that trustworthiness can be improved by a trust-oriented interface design which aims at improving trust related attributes. As it is unknown what influences public trust in Web GIS, the wider research framework that this study follows is based on an investigation of different interfaces with non-expert users using HCI methodology in order to understand how trust perceptions are formed. In particular this paper investigates the applicability of trust related attributes described in the literature, using the WIYBY website, which has a GIS element and also incorporates the elements of uncertainty, risk and complexity.

## 2. Methodology

Chopra and Wallace (2003) define online trust as a person's (trustor) willingness to depend or rely on an online system (trustee) (Figure 1). In this relationship, the trustee attributes are of particular importance, which can facilitate the design process of a more trustworthy system.

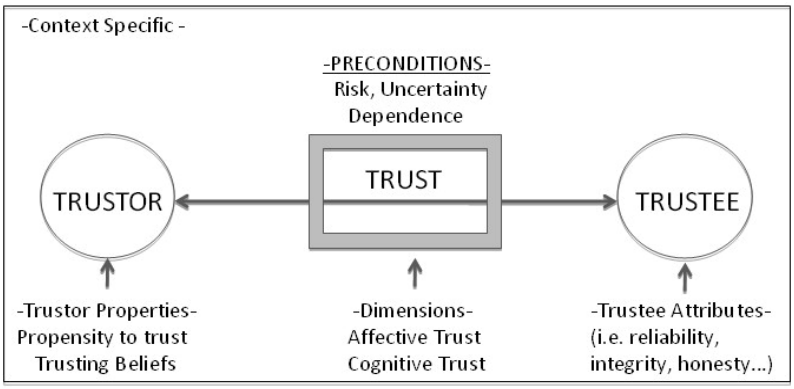



**Figure 1:** Trust Components.

Several studies attempt to describe the trust related attributes (e.g. Corritore et al., 2003) and it can be summarised that these fall within two categories of attributes. The perceptual attributes concern the source and its reputation and trust cues such as logos, testimonials and blogs can increase trustworthiness. The functional attributes involve evidence collected through interaction and assessment of the system's quality. In this category aesthetics, usability and a good user experience can improve trustworthiness.

It is unknown whether these trust related attributes have the same influence in trust perceptions in the Web GIS context. For this reason the aim of this paper is to understand the effect of these attributes (aesthetics, usability and user experience) and the role of trust cues in the perceived trustworthiness of the WIYBY website.

HCI methodology, which examines how people use computers, can support this investigation and Usability Testing (UT) was the technique used, together with pre and post-test questionnaires and Think Aloud (TA) data of user comments. As the study focuses on non-expert users, the pre-test questionnaire used to gather user data to ensure that nobody used the WIYBY website and had no GIS expertise. The post-test questionnaire measured the constructs under investigation using a Likert scale from 1 "Strongly Disagree" to 5 "Strongly Agree". The initial questionnaires used duplicates and double negatives to ensure that users focus on their selections and for the presented herein results some questions, including GIS specific, were removed to ensure internal consistency.

The usability construct was based on nine items of well-established usability heuristics and previous studies (Flavian, 2005). It should be noted that the GIS element is central and thus the items refer to the overall usability, including the GIS interface. One open question asked the users to describe any specific GIS related difficulties. The aesthetics construct was measured based on five items using some of Nivala et al. (2008) GIS usability principles which focus on aesthetics. The UXP construct was created based on the characteristics of a good UXP (e.g. the application should be engaging, not stressful, and pleasant).

The trust cues element created based on trust cues suggested in the literature (e.g. Cheskin, 1999) and the users were asked to consider their applicability in the WIYBY and helped to understand trust related user expectations, which never examined for Web GIS.

The trustworthiness scale was the most problematic as existing studies focus mainly on e-commerce and therefore many items are irrelevant to Web GIS. The final trust scale was based on four items

following Fogg and Tseng (1999) suggestions for credibility assessment. Additional open questions were used for all constructs.

## 3. Case Study and Results

The WIYBY website[1] was examined using six tasks in July 2009. Ten non-experts were recruited, as previous studies showed that five users can identify the majority of usability problems (Nielsen, 1994). During the usability testing the actual usability of the website was measured, using the success rate formula (Nielsen, 2001), and was 60.8%, although each user failed in at least one task.

The perceived usability, which describes what users think about the WIYBY usability, was low (Figure 2), with 7 out of 10 users disagree that it was easy to use and 8 users think that it needs improvement. It should be noted that for Figures 2-6, the "Agree" and "Strongly Agree" answers were grouped and the same was done for the "Disagree" and "Strongly Disagree" answers. The mode which is the most frequent answer is provided, as well as, Cronbach Alpha to assess reliability of data.

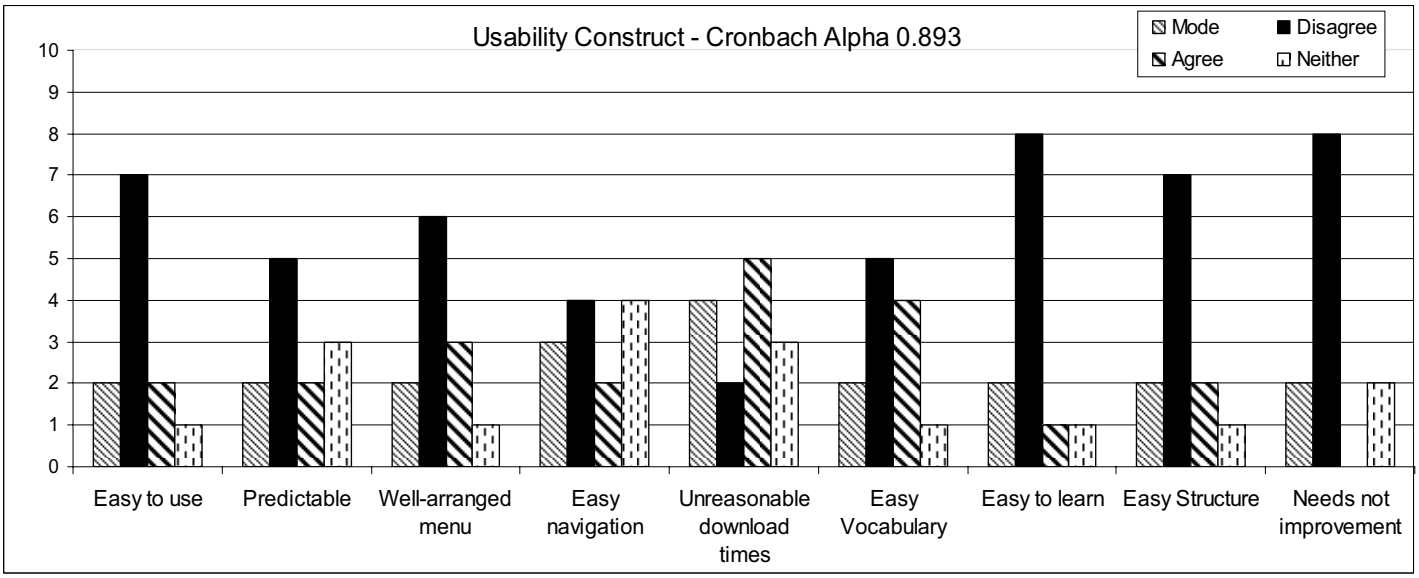



**Figure 2:** Usability construct: post-test questionnaire.

The aesthetics were also problematic (Figure 3). All users agreed that map visualisation needs improvement, and the GIS element was not considered as usable, because of overlapping results, poorly designed maps and not clear symbology.

Although previous studies suggest that usability and aesthetics influence trust perceptions, these had a limited influence in the system's trustworthiness. The trust construct shows that 9/10 users trusted all the information provided, although the majority found it not transparent. Based on open ended questions and Figure 4, the main reason for trusting the information was because it is provided by EA. Think aloud data show that transparency influenced by usability problems and its poor design.

[1]URL: http://maps.environment-agency.gov.uk/wiyby/wiybyController?ep=maptopics&lang=_e

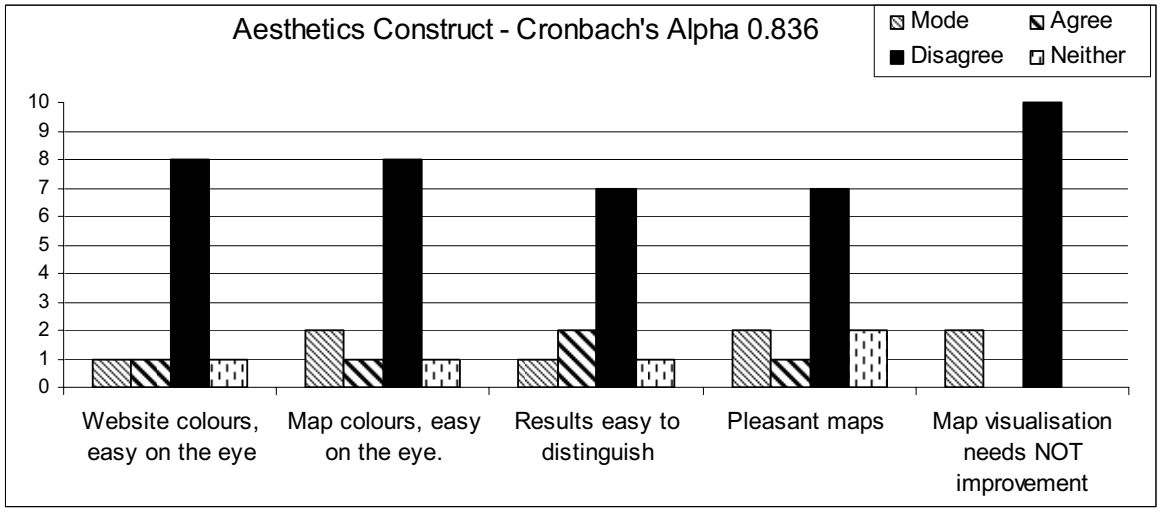



**Figure** 3**:** Aesthetics construct: post-test questionnaire.

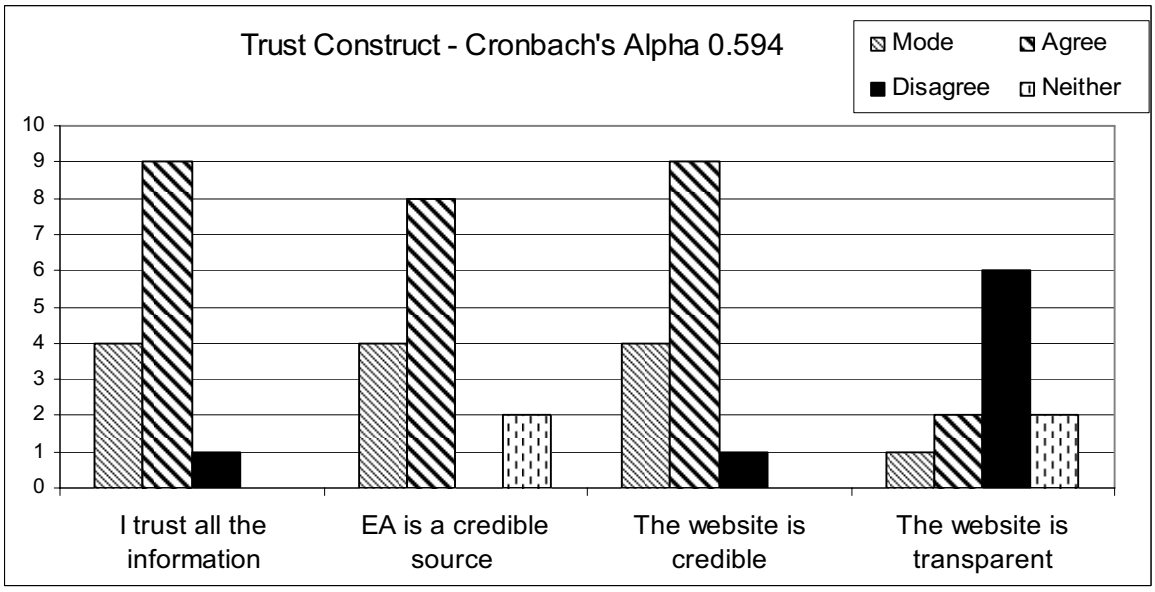



**Figure** 4**:** Trust construct: post-test questionnaire.

While most users trusted the information, the UXP was problematic and 8 users will not use the website again (Figure 5). Affective trust creates a bond with the system, which means that for the WIYBY website, trust formations were based on cognition (cognitive trust), and mainly the source's reputation, which is a cognitive construct.

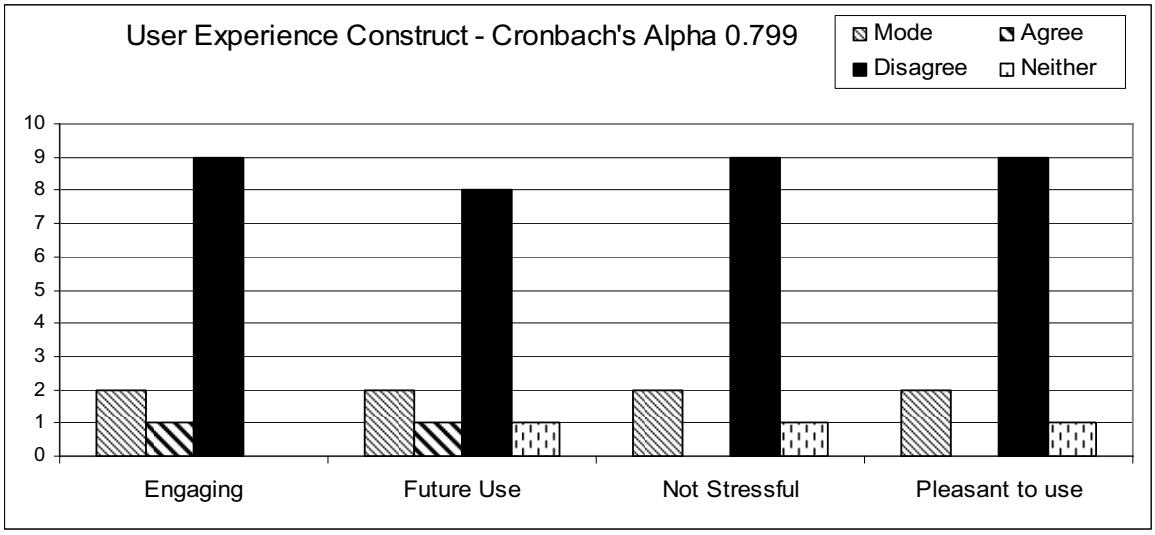



**Figure** 5**:** User experience construct: post-test questionnaire.

Figure 6 provides an overview of the users' opinion about trust-inducing features, described in the literature.

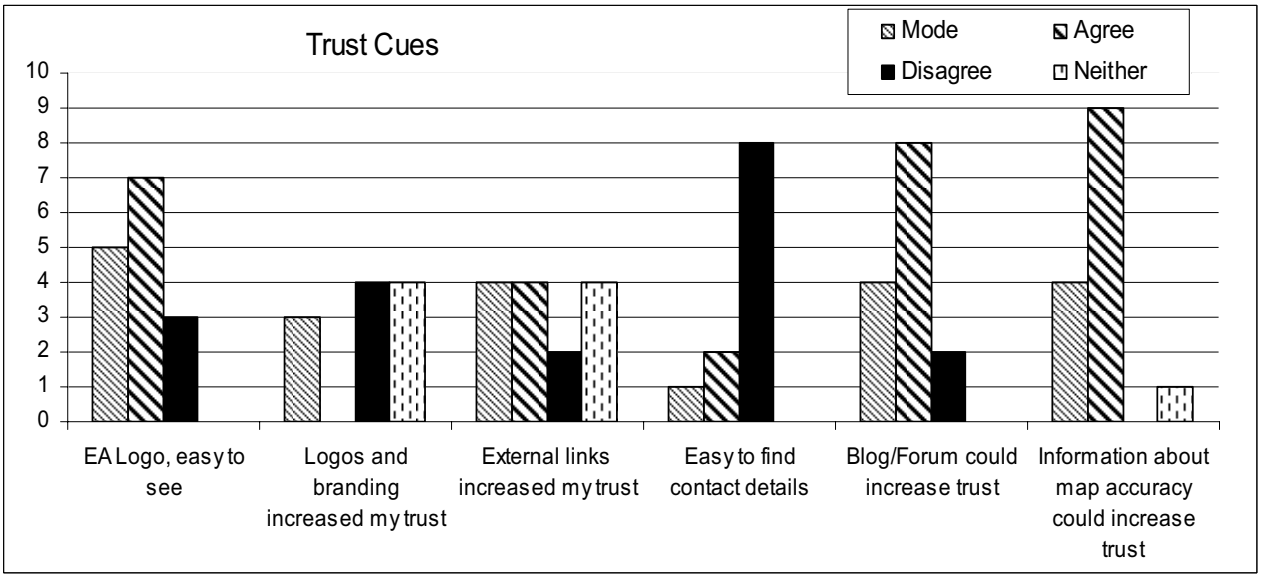



**Figure 6:** Summary of users' answers about trust-cues.

## 4. Discussion and Conclusions

The results of this study show that usability, aesthetics and UXP were problematic. Also, the majority of users trust the information provided although they do not think the website is transparent. Figure 4 and qualitative data explain that it is the source's reputation that had a major influence in trust perceptions (e.g. User 7: *"I would trust information coming from a Government Agency on principle"*).

Although in this study the functionality dimension was known and weak, the users still trusted the information. It implies that people rely on reputation although they recognise problematic aspects and this is another paradox of online users. This finding is in line with *'halo effect'* described by Fogg (2003, p.124). However, if the functionality influences the trustworthiness of a less reputable source needs further investigation.

The study showed that users had difficulties with the GIS element and that the navigation and overall structure were confusing. The perceived usability and aesthetics influence the perceptions about future engagement. Most users felt they will not use the system in the future (i.e. User 5: *"Not particularly easy to use, would probably rather use an alternative information source"*). This indicates that more emphasis should be paid by designers not only in the actual but also in the perceived usability.

The map visualisation was problematic for mainly aesthetical reasons but also prevented the users from understanding the information, although no comments of suspicion were made for map-based information. However, users suggested that information about maps' accuracy and construction could increase their confidence and trust.

This study is a preliminary investigation of trust related attributes, described in the literature, in the Web GIS context. It should not be ignored that Web GIS have their special user aspects and thus additional HCI methods such as Cooperative Discovery, which allow evaluators to interact with users while using the system, could provide a deeper insight on additional attributes that influence trust. To better understand how trust perceptions are formed, different Web GIS platforms should be investigated. Also this study focuses on non-expert users, who have different needs from experts, but experiments with expert users could reveal differences in the formation of their trust perceptions.

## 5. Acknowledgements

This project is funded by the EPSRC Engineering Doctorate of Computer Science at University College and Arup Ltd.

## References

Chopra K and Wallace WA (2003) Trust in Electronic Environments *Proceedings of the 36th Annual Hawaii international Conference on System Sciences,* Hawaii January 06-09, pp 331-340

Cheskin Research and Studio Archetype/Sapient (1999) eCommerce trust study [Online] Available at: http://www.cheskin.com/cms/files/i/articles//17__report-eComm%20Trust1999.pdf [Accessed 10 May 2009]

Corritore C L, Kracher B and Wiedenbeck S (2003) On-line trust: concepts, evolving themes, a model *International Journal of Human-Computer Studies* **58** pp 737-758

Flavian C, Guinaliu M and Gurrea R (2005) The role played by perceived usability, satisfaction and consumer trust on website loyalty *Information & Management,* **43(1)** pp 1-14

Fogg BJ (2003) *Persuasive Technology: Using Computers to Change What We Think and Do.* Morgan Kaufmann Publishers, San Francisco, CA

Fogg BJ and Tseng H (1999) The Elements of Computer Credibility *Proceedings of CHI'99 Human Factors in Computing System,* Pittsburgh May 15-20, pp 80-87.

Haklay M and Zafiri A (2008) Usability Engineering for GIS: Learning From A Screenshot *The Cartographic Journal* **45(2)** pp 87-97

Nielsen J (1994) Guerrilla HCI: using discount usability engineering to penetrate the intimidation barrier. In Bias RG and Mayhew DJ *Cost-Justifying Usability*. Academic Press, Forida

Nielsen J (2001) Success Rate: The simplest usability Metric [Online] Available at: http://www.useit.com/alertbox/20010218.html [Accessed 25 November 2009]

Nivala AM, Brewster S and Sarjakoski LT (2008) Usability Evaluation on Web Mapping Sites *The Cartographic Journal*, **45** pp. 129-138

Schneiderman B (2000) Designing trust into online experiences *Communication of the ACM* **43(12)** pp 57-59

Skarlatidou A and Haklay M (2006) Public Web-Mapping: Preliminary Usability Evaluation *Proceedings of GIS Research UK 2005*, Nottingham April 5-7

**Biographies**

*Artemis Skarlatidou, is an EngD student at University College London (UCL), where she also completed her MSc in Geographical Information Science. Her main research interests concentrate on HCI aspects of GIS and web-based GIS applications. In her EngD project investigates the concept of trust in Web GIS in order to improve the trustworthiness of these systems through interface design.*

*Mordechai (Muki) Haklay is Senior Lecturer in Geographical Information Science at the Department of Civil, Environmental and Geomatic Engineering, UCL. He has been working on socio-economic analysis using GIS, novel analyses techniques with GIS and usability aspects of GIS. He has published papers on these subjects in Area, International Journal of Geographical Information Science, and the Journal of Environmental Management.*

*Tao Cheng is a senior lecturer in GeoInformatics in the University College London. She has studied and lectured in China, the Netherlands, Hong Kong, France and the UK. Tao Cheng's research spans spatial-temporal modelling, analysis and visualisation, uncertainty and quality of geographic information, and spatio-temporal data mining and knowledge discovery. She has over 90 publications and she received The U. V. Helava Award for the Best Paper in the ISPRS Journal of Phogrammetry and Remote Sensing in the year 2000.*

*Nicola Francis completed her BSc in Physical Geography at the University of Durham and her MSc in Surveying at UCL. Her MSc Thesis was woncerned with investigating the usability and trustworthiness of certain Web GIS page, with particular reference to eye-tracking technique.*

# Taking GIS out of the classroom: developing effective learning environments with mobile GIS

Kenneth Field[1], James O'Brien[1]

[1]Kingston University London, Centre for GIS, Kingston upon Thames, London KT1 2EE
Tel. +44 (0)20 8417 2541
Email Kenneth.field@kingston.ac.uk
http://www.kingston.ac.uk/centreforgis

KEYWORDS: mobile GIS, data capture, fieldwork

## 1. Introduction

Fieldwork has long been the centrepiece of experiential learning in geosciences (Coorey, 1992; Kolb, 1984) and gives students vital mechanisms to deploy and develop skills and enhance understanding in the real world. Students use the real-world laboratory to study patterns and processes and to enhance their understanding of a wide range of the subject matter. "The dominant style of fieldwork which has developed is the excursion-type, commonly called the "Cook's Tour" characterized by a didactic/instructive teaching approach with passive student interaction" (Hawley, 1996, p243). Clearly, such an approach fails to grasp the opportunity presented by fieldwork to develop key skills in an exciting and dynamic manner (Grattan et al. 2005).

Kingston University's Centre for GIS places a strong focus on fieldwork. With the undergraduate GIS degree being the longest running in the world the teaching team strives to develop realistic fieldwork experiences for students across all years of the program that takes GIS out of the classroom and which harness developments in mobile mapping and GIS (Clarke, 2004)

First year GIS students undertake a fieldcourse on Isle of Wight alongside other geoscience students and second year students undertake a week-long mobile GIS fieldcourse to Malta. Both fieldcourses support the development of data collection and mobile mapping techniques with the latter taking a more critical approach with a strong emphasis on collaborative learning and reflection. To accomplish these tasks students utilize a combination of ArcPad and ArcGIS Mobile to rapidly collect data in the field, share it on the fly with each other and teaching staff and collaborate to build a shared understanding.

This paper outlines the development of efficient mobile mapping environments to support student learning, with a focus on the Malta fieldcourse, and describes the ways in which the techniques are deployed. Such developments in implementing mobile technologies in fieldwork are important in enhancing field based learning and the wider curricula.

## 2. Background

The challenge to be addressed by the project was to develop new processes and workflows that harness the developments in mobile GIS (Tsou , 2004) to replace many of the bottlenecks in current practice. Specifically, locally installed mapping packages (e.g. ArcPad) running on PDAs or laptops (Wagtendonk and Jeu, 2007)mean the dominant activity is data is gathered during fieldwork and then

combined at a later stage in a desktop environment. This creates significant delays in the field and can create problems if students have not collected data in the same way as another group for instance. The fostering of a collaborative approach to data gathering is only partially served by this type of approach (Drummond *et al.* 2006) and limitations exist in terms of what different groups of students achieve. For instance, groups tend to operate independently from each other and gather data in different ways (e.g. classifying features using different object types, capture resolutions, attribution and detail) which creates problems when data is combined. Results are often disappointing and have done little to enhance more traditional data collection for communal mapping.

This project explored the potential for creating local area networks in a fieldwork environment using mobile wireless routers and mobile broadband and to have students collect data and serve it back to a central hub on-the-fly. The benefits of such an approach are the rapid collation of a central database which will be ready for dissemination to students on return. There is also potential for students to have updates streamed to their own devices to illustrate what data other groups might be capturing; thus avoiding repetition and to share information on the scale of data capture, the resolution, landscape type and feature recognition. The project aimed to deliver improvements in workflow and group interaction by testing the application of these new methods.

There are a number of benefits that this approach to understanding GIS work in mobile environments might have in enhancing the student experience and their spatial literacy:

- creating mechanisms for student-student group interaction when small groups are working in isolation across wider study areas as part of a larger exercise; providing a means to collaborate remotely in order to bring together interim results or develop methodologies as part of the exercise itself.
- providing a means of enhancing student-staff interaction for the purposes of support; this could be to assess interim pilot work for formative feedback in remote locations or the modification of data and data requirements delivered by staff who can monitor streamed data in the server and respond accordingly.
- develop content that can be delivered remotely in the field to enhance student-mental interaction (e.g. to provide challenges during an exercise or to introduce changes to data requirements to assess adaptability and ingenuity)
- disseminating best practice and methodologies based on the work to date such that other institutions may benefit in developing their own curricula.

## 3. Method

### 3.1 Pre-fieldcourse development

To manage mobile fieldwork in Malta an ArcGIS Server (v. 9.3.1) was setup at Kingston University. This server running on Windows 2003 Server was populated with orthophotos and map extracts as a frame of reference for students. Data was published to the server using the standard ESRI ArcGIS and ArcCatalog workflows which will be outlined here very briefly. A Microsoft SQL SDE database was created on the server and populated using ArcCatalog with shapefiles and imagery. The data was then added to ArcMap, trimmed to the boundaries of the study area and saved back to the SDE database. This data is then published to the ArcGIS server using the ArcPad toolbar. The map document produced is then added as a Map Service (in ArcCatalog) to the ArcGIS Server. The map service is created and any data are uploaded to the ArcGIS Server.

The map service created will be visible within the ArcGIS Server Manager console (accessed from a web browser) enabling data preview, metadata editing and web mapping application creation. A web mapping application is then created on the server and layers added to it from the map service created earlier in ArcCatalog. Once the web application is created and started it can be accessed from any web browser and the map data used (Figure 1).


**Figure 1.** Example fieldcourse map service in web browser

The same data contained in the map document was also copied to a set of Trimble Juno SB PDAs (to reduce data download time and costs in the field). The data can be downloaded over WIFI or 3G to the devices though if required. Once in the field students could add and edit data using ArcPad running on the Juno. Data could be synchronised back to the Kingston server using either a 3G/GPRS connection (via a paired mobile telephone) or a local WIFI network with an internet connection (via a mobile broadband dongle). For the Malta field course this server was running within a virtual machine on a laptop to make support and maintenance easier (Figure 2).

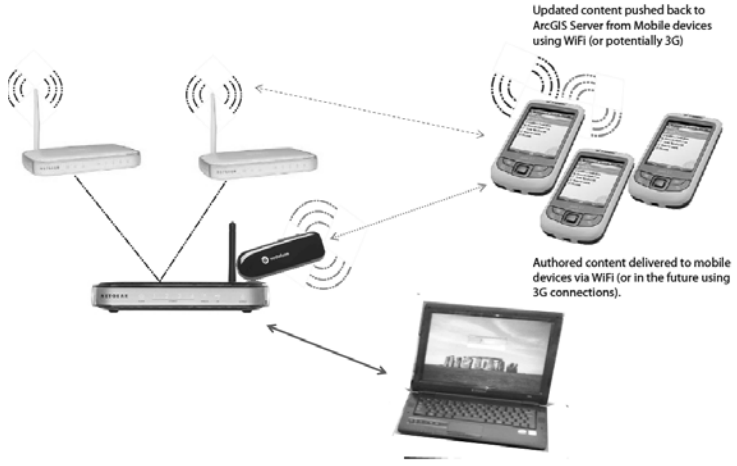



**Figure 2.** Deploying ArcGIS Mobile for fieldwork

The synchronisation process allowed students to share their data with students but also to receive updates from students working in other areas of the study area. Data uploaded to the server is versioned so that multiple versions of the same item can be managed if necessary. Staff monitoring student uploads were able to track student locations and progress during data collection.

### 3.2 Fieldcourse delivery

An initial service area analysis is defined to create six group study areas around the main study area in Mellieha in northern Malta (Figure 3). Students then have a day in the field to create a comprehensive topographic survey and land use map. The initial briefing is deliberately limited in detail to encourage a collaborative approach and improved performance when in the field.


**Figure 3.** Mellieha study area in Northern Malta

Small groups of students use the Trimble Juno SB with an internal GPS running ArcPad. A digital orthophoto backdrop provides students with spatial reference information and a customized ArcPad interface allows polygon, line, point and attribute input. The data layers are saved on the device before being uploaded to the server via wireless broadband or WIFI.

The ArcGIS server contains data from previous surveys that students can download to their PDAs to perform rudimentary land use change calculations. Academic staff monitor the students' uploads and advise via other channels (e.g. txtools) as they work towards developing a common, shared understanding of the various issues surrounding the feature recognition and representation on-the-fly. The students can also view other groups' land use or risk classifications to ensure they are working towards a consistent map that builds upon shared understanding.

One of the first issues to overcome is the recognition of different, unfamiliar, landscape types. For instance, the pre-defined attribute fields in the ArcPad data layers have land uses such as 'Maquis' and 'Garrigue'. Students soon find themselves unable to interpret the meaning of the term and associate it with a feature. Using txtools and a Twitter Map, students can communicate with each other and with staff to discuss the feature type and, using Twitpic and data that is delivered to their PDAs, it is possible to send images of the different features. In this way, a common, shared understanding is developed that improves the consistency of data collection between groups. The process also illustrates the importance of ensuring appropriate classifications and understanding before embarking upon such an exercise.

Students create a range of products from the data collected including basic topographic maps illustrating the decline in agriculture as overlays (Figure 4), 3d visualisations of the landscape and surveyed data (Figure 5) and different scales of land use mapping (Figure 6 and 7)

(a) (b)

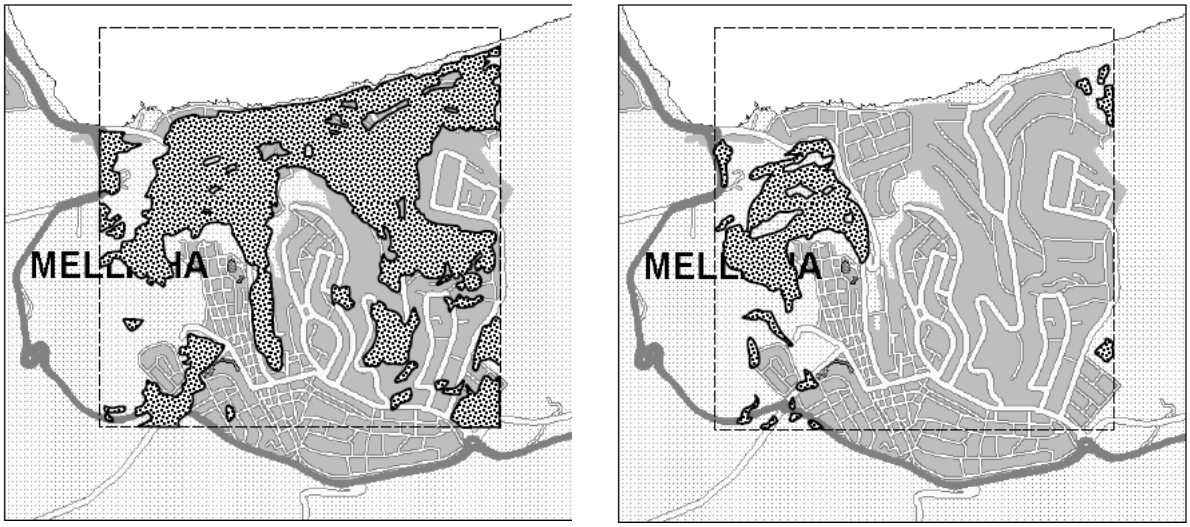


**Figure 4.** (a) agricultural decline in Mellieha 1956 survey
(b) agricultural land use in Mellieha 2009 survey

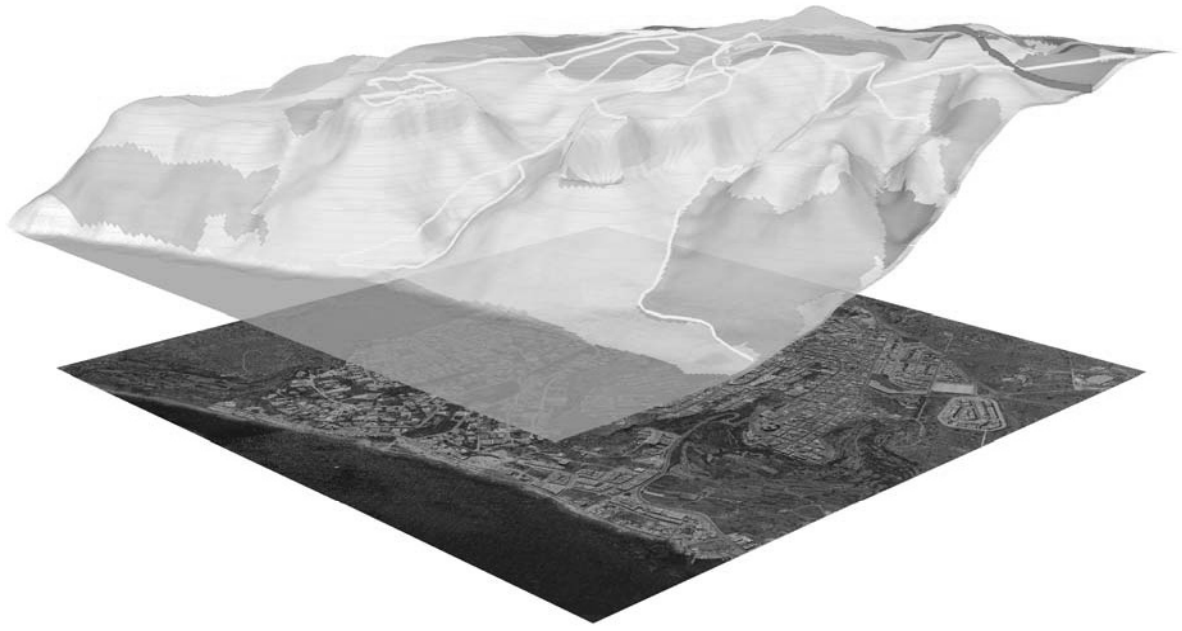


**Figure 5.** Surveyed landuse data drape on DEM above digital orthophoto, Mellieha 2009

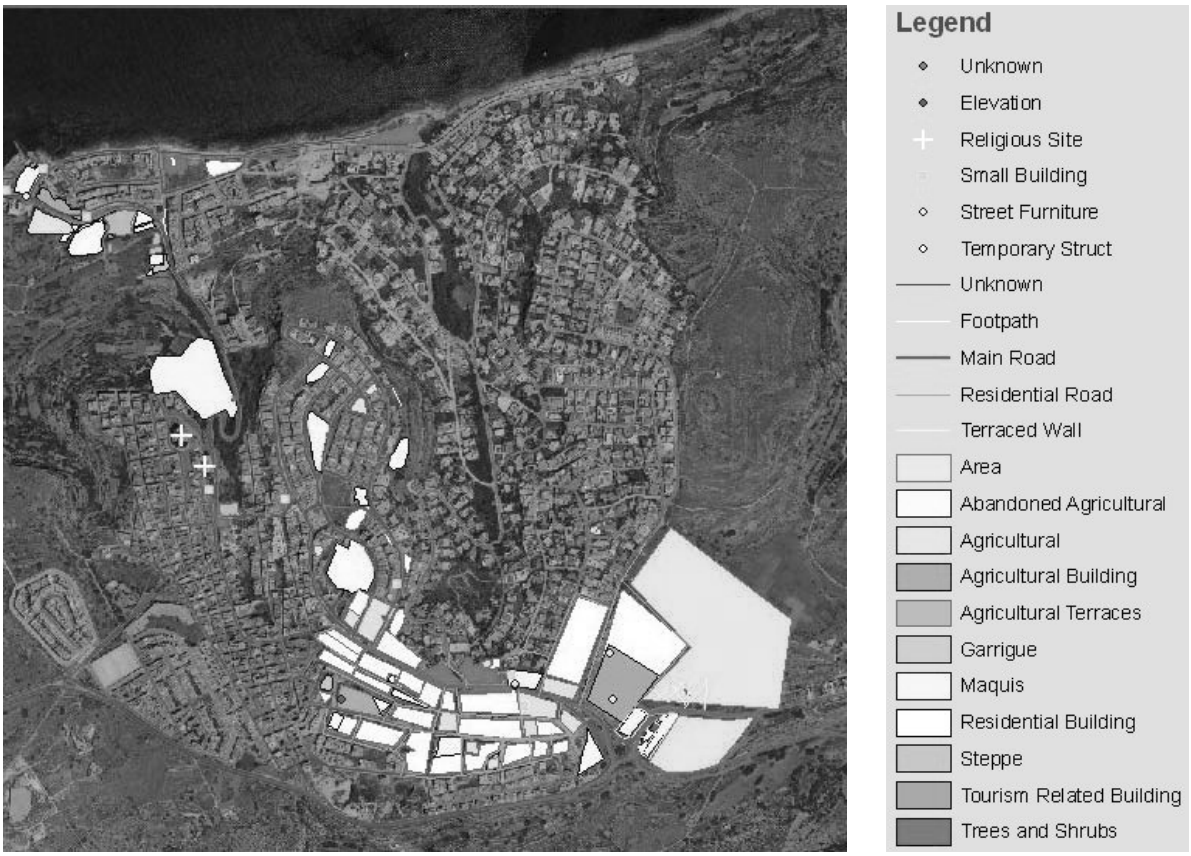



**Figure 6**. Medium-scale land use survey from one group, Mellieha 2009

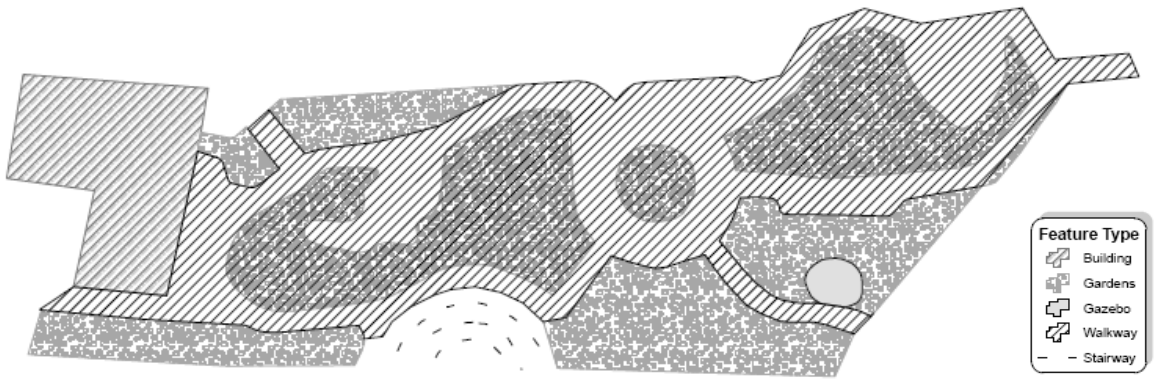



**Figure 7**. Large-scale topographic detail from one group, Mellieha 2009

By comparison to previous fieldcourses a number of benefits to the approach were indentified. Previous work has involved visual surveys, field sketching, rough calculations, smaller study areas and poor hand-drawn maps. Even when we developed our initial mobile mapping capabilities in 2005 the technology was not sufficiently mature to support learning, with many pitfalls associated with battery life, rudimentary equipment and software which was insufficient for the task at hand. Barriers have now diminished with reducing costs of equipment, increased flexibility & usability, improvements in wireless & integrated technology, increased synergy of geospatial technology and realistic implementation.

Since we have incrementally improved our mobile GIS work the focus has been less on making the technology work, to focussing on the tasks. Students have previously received considerable preparation for feature recognition, object classification and attribution but technology is now stable enough that we can focus on less pre-trip preparation (that is largely forgotten) and the development of learning through implementation. The results of previous maps have not been impressive as patchy recall has led to less than optimum data collection. The absence of prior information required students to be proactive in the field to identify problems and to interactively learn from one another via the building of the map as well as through collaboration using a range of tools (e.g. a Twitter Map and the use of txtools).

There are few drawbacks to the work, particularly since hardware has developed to an extent that it is much more dependable than even a few years ago. Battery life is sufficient though we do supply spare batteries. The only operational drawback in Malta is the screen glare caused by the reflected sunlight. A number of pedagogic benefits exist which can be summarised as follows, based on reviews of exercises through focus groups with students and assessments of the quality of work year on year:

- mechanism to introduce & implement new technologies;
- development of core competencies in data collection;
- achieve a greater engagement by students;
- integrate specialist with ubiquitous technology;
- extends key skills and develops critical thinking in key geospatial areas;
- improved depth, breadth & quality of data;
- increased scope for analysis, interpretation & visualisation;
- allows real world consideration of benefits & limitations of approaches, error, accuracy, uncertainty; and
- improved map output.

Taking GIS education into the outside world provides context and direct links to the real world. It enables the understanding of the difference between real-world & modelled environments and improves understanding of how to represent the real world. It gives students an appreciation of where data comes from (rather than reliance on secondary, derived datasets) and an understanding of error & uncertainty introduced when translating from the real-world to modelled world. Awareness of the need to question data, data quality & data structures can be more readily achieved as well as developing a first hand understanding of metadata & metadata coding. Fundamentally, mobile GIS fieldwork provides the capability to understand the analytical framework – showing how interpretations of a modelled environment relate back to the real-world.

## 4. Outcomes

The methodologies and workflow created to support the land use mapping project in Malta have

improved the exercise and student experience considerably. It has allowed the introduction of debates and collaborative approaches to developing common, shared frameworks for understanding in the field through interaction via a range of media. The mapped results have allowed much more rapid gathering of higher quality and more consistent data as well as enabling students to realise the full value of mobile GIS environments for data capture. Hardware and software is now stable enough to perform the required tasks and our workflows enable efficient implementation. We extend beyond replicating lab-based work, enhance and develop conventional fieldwork and give GIS students opportunity to 'get their hands dirty'

## 5. Acknowledgements

This work is part funded by an Honorary Visiting Fellowship in the SPLINT CETL at the University of Leicester from 1 July 2009 to 31 March 2010.

## References

Clarke, K. (2004) Mobile mapping and Geographic Information Systems, *Cartography and Geographic Information Science* **31(3)** pp.131-136

Coorey, P.G. (1992) Fieldwork : an essential component of geoscience education and training, Episodes, 14 (2) pp337-340

Drummond, J., Billen, R., Joao, E. and Forrest, D. (eds) (2006) Dynamic and Mobile GIS: Investigating changes in space and time (Innovations in GIS), CRC Press, London

Grattan, J., Gilbertson, D. D. and Horgan, J. (2005) Skills Development via Fieldwork: The Malta Experience, UK Geosciences Fieldwork Symposium: Proceedings (Online: http://www.gees.ac.uk/essd/field.htm#GrattanJ) Accessed January 2010

Hawley, D. (1996) Changing approaches to teaching Earth Science fieldwork. Geoscience Education and Training. 19, 243-253.

Kolb, D. (1984) Experiential learning: experience as the source of learning and development. London. Prentice Hall.

Tsou, M-H. (2004) Integrated Mobile GIS and Wireless Internet Map Servers for Environmental Monitoring and Management, *Cartography and Geographic Information Science* **31(3)** pp.153-165

Wagtendonk, A. J., and Jeu, A. M. (2007) Sensible field computing: Evaluating the use of Mobile GIS methods in scientific fieldwork, *Photogrammetric Engineering and Remote Sensing* **73(6)** pp.651–662

## Biography

Dr Kenneth Field is Principal Lecturer in GIS and Course Director of undergraduate and postgraduate GIS courses at Kingston University. He is current Editor of The Cartographic Journal with particular interests in cartography, geovisualization and mobile GIS. He is currently in receipt of an Honorary Visiting SPLINT CETL Fellowship.

Dr James O'Brien is Principal Lecturer in GIS at Kingston University. His particular interests are in spatial databases, GIS software development, Semantic networks and ontologies and the application of mobile GIS for collaborative data capture.

# Layout and Colour Transformations for Visualising OAC Data

Jo Wood, Aidan Slingsby Jason Dykes

giCentre, Department of Information Science,
City University London, Northampton Square, London, EC1V 0HB, UK
Tel. +44 (0)20 7040 0180
sbbb717|jwo|jad7@soi.city.ac.uk, http://gicentre.org
Colour version of this paper available at http://gicentre.org/papers/gisruk10/wood_layout_2010.pdf

KEYWORDS: OAC, cartogram, mapping, perceptually uniform colour

## 1. Introduction

The Output Area Classification (OAC) is a geodemographic classifier that characterises the UK population by socioeconomic characteristics at Output Area (OA) (Vickers and Rees, 2007). Categories are based on clusters identified by *K*-means clustering of 41 of the 2001 census variables. Table 1 shows the 7 OAC super-groups along with some of their characteristic census variables. Each Output Area is allocated its most similar OAC category resulting in significant data generalisation. This is both intentional and essential to their utility. Knowledge of such variability between and within categories is beneficial to its use.

**Table 1.** OAC super-groups with the census variables that deviate most from the national mean, using our new colour scheme with the SASI (2009) colours on the left.

| | |
|---|---|
| Blue collar communities | **High:** terraced housing<br>**Low:** flats, HE qualifications. |
| City Living | **High:** HE qualification, live alone, born outside UK, rented (private), flats<br>**Low:** detached housing, have non-dependent children, 5-14 (age). |
| Countryside | **High:** 2+ cars, work from home, agriculture/fishing, detached housing<br>**Low:** population density, take public transport, flats |
| Prospering Suburbs | **High:** 2+ cars, detached housing<br>**Low:** public rent, terraced housing, flats, no central heating, rented |
| Constrained by Circumstances | **High:** flats, rented (public)<br>**Low:** detached housing, 2+ cars, HE qualifications. |
| Typical Traits | **High:** terraced housing<br>**Low:** rent (public) |
| Multicultural | **High:** Ethnic minory, born outside UK, flats, rented, take public transport<br>**Low:** detached housing |

The use of OAC in both local and national governments for analysis, for reporting statistics and for studying the impact of policies is increasingly encouraged (e.g. DCLG, 2009, p.73) and graphical techniques for depicting such data are becoming increasingly important (e.g. LCC, 2008).

We present a national map of the 259,847 OAs in England, Scotland and Wales coloured by their closest super-group and degree of typicality. We also provide separate maps for each super-group. In order to present these data effectively, we use a hierarchical cartogram as a base map that normalises area by population density and a colour scheme in which hue indicates the super-group and lightness indicates its degree of typicality. The contributions of this work are:

- Space-filling rectangular hierarchical cartograms for efficient depiction of all output areas.
- A new colour scheme that allows comparison of super group categories and classification uncertainty.

The cartograms can be used with any spatial hierarchy, are scalable (we have successfully mapped all the >1.7 million UK postcode units in their postcode hierarchy) and can depict two data variables simultaneously using size and colour (when generalised to treemaps, position can be used to depict further information; Slingsby *et al*, 2009).

## 2. Colour Selection

Perceptual variation in hue, saturation and lightness is non-linear. This has strong implications for the design of colour schemes (Brewer, 1994). The colours used by SASI (2009) (left of Table 1 and Figure 2) vary in hue, lightness and saturation making the red and yellow more prominent than grey, giving them unintentional and undue emphasis. Perceptual colour spaces can help us design colour schemes that allow categories to be compared more effectively. We use the CIELuv colour space as recommended by Wijffelaars *et al* (2008) in which distance between colours in this colour space is proportional to perceptual discrimination. Figure 1 graphs the distance between hues at 5-degree intervals on a hue colour wheel in the CIELuv space at three different lightness values, where higher values indicate better perceptual discrimination at that hue. This shows that the perceptual discrimination between hues is variable and varies greatly with lightness (at 30% lightness, red/orange is particularly distinctive; at 70% lightness, the level of distinction is lower and is strongest for yellow/green).

**Figure 1:** Graphs showing the ease of hue discrimination at lightnesses of 30%, 50% and 70% (left to right). Distance from the centre of each graph is proportional to degree of perceptual discrimination of the given hue.

In depicting typicality with lightness, it is more important to distinguish hues at the darker end of the lightness scale. We therefore identified seven hues at perceptually equal intervals for a lightness of 30%, and then allocated them to the OAC categories to match the SASI (2009) scheme as closely as possible (see Table 1). Colours are less discriminatory since saturation and lightness are equal, have equal perceptual prominence and are perceptually equally spaced. Since lightness is held constant, we can map it to another variable. As its perception is non-linear and hue-dependent, we use the CIELuv colour space using Wijffelaars *et al*'s (2008) method to generate colour palettes for each hue where lightness varies in a perceptually-linear manner.

In Figures 3 and 4, lightness indicates the degree of typicality ($T_g$) to a super-group ($g$):

$$T_g = 1 - \frac{d_g}{\max(d)} \quad (1)$$

where $d_g$ is the distance to cluster $g$, max($d$) is the distance to the furthest cluster and $d_g$=min($d$) when the nearest cluster is used.

### 3 Space-filling hierarchical rectangular cartograms

Figure 2 is a map of OAs produced by SASI (2009) coloured by OAC super-group. Since each OA each contains a similar number of households (50-100), sizes vary greatly depending on population density. This makes their depiction on national maps challenging. 'Countryside' (green) dominates the map when, in fact, it relates to a very small proportion of the population. Major urban areas are visible as fine-grained heterogeneous patches, but the structure of these is impossible to detect.

Density-normalising cartograms are widely used for this type of problem (e.g. Dorling et al., 2008). They have the effect of enlarging the parts of the map with high population density at the expense of those parts with low population density. This causes necessary geometrical distortion, the nature of which depends on the type of cartogram. Gastner cartograms (Gastner and Newman, 2004) distort shape but maintain contiguity, whereas rectangular (Florisson et al., 2005; Wood and Dykes, 2008) and circular (Dorling, 1996) cartograms fix shape to assist in size comparison, at the expense of area contiguity.

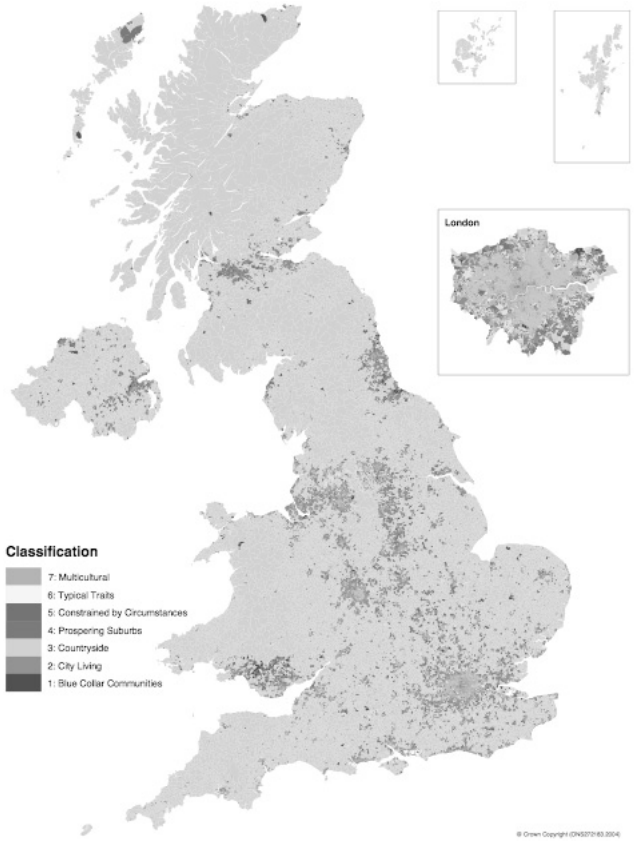



**Figure 2.** OAC super-groups mapped to Output Areas using SASI colours (SASI, 2009)

In Figure 3, we use a space-filling rectangular hierarchical cartogram that displays OAs as rectangles sized by population, using the colour scheme described in section 2. An advantage of space-filling cartograms is their efficient use of the space through the complete tessellation of the data into a rectangle. The corresponding disadvantage is that contiguity between areas is not necessarily preserved and there may be considerable positional displacement. However, the hierarchical nature of this cartogram allows us to map the OAs within the UK postcode hierarchy, a well-recognised spatial frame of reference.

Although the some of the geographical contiguity and positioning is lost (leading to the linear artefacts around the edge of postcode areas), the broad structure of the geography is preserved. Purples, greens and yellows are either more prevalent or spatially autocorrelated and the broad geographic differences within the 'multicultural' class are evident.

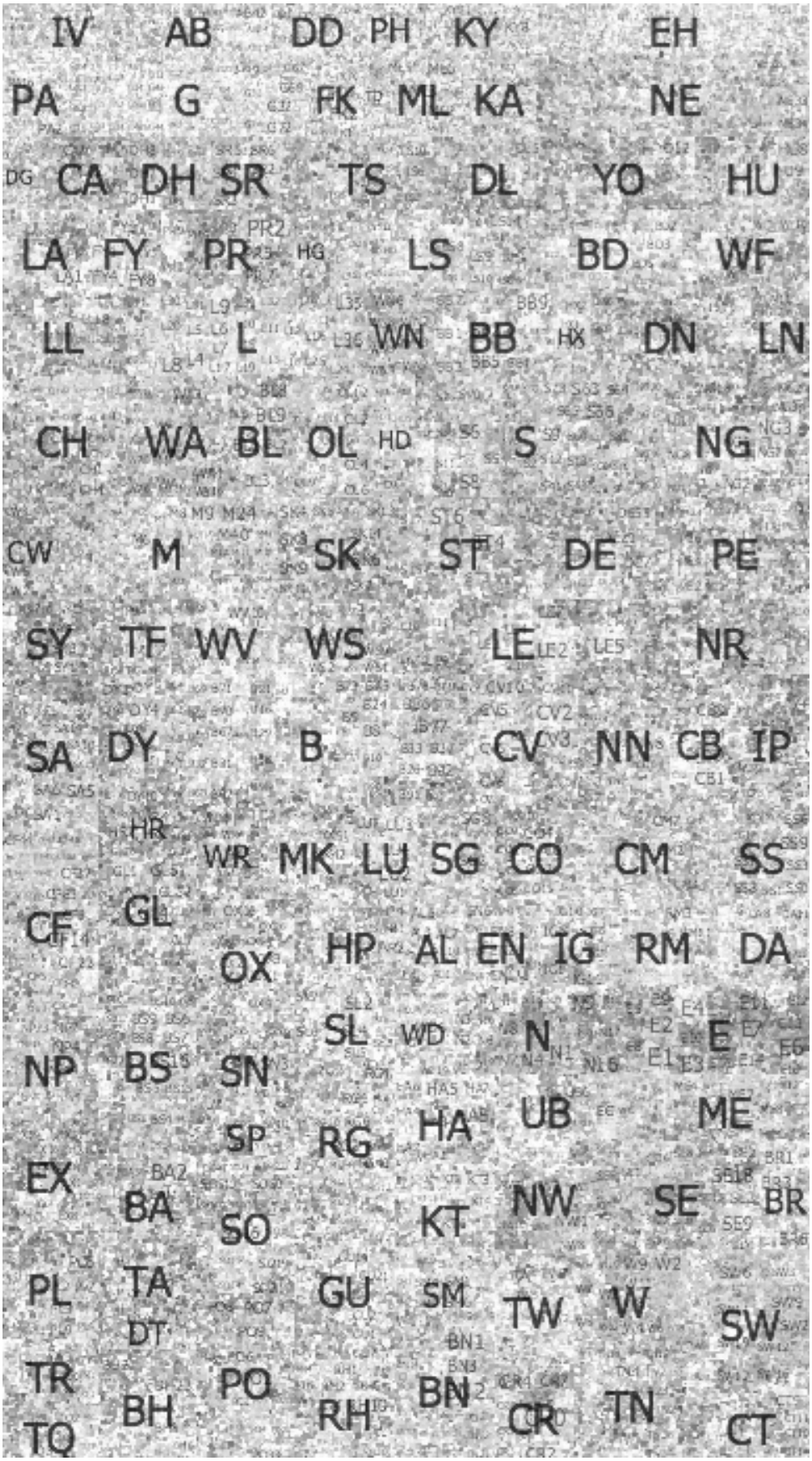



**Figure 3.** Output Areas sized by population within postcode using a rectangular hierarchical cartogram. Hue indicates OAC super-group (Table 1) and lightness indicates atypicality.

## 4. Interpretation

The interpretation of Figure 4 is much easier as a large high-resolution poster, however many clear and interesting patterns are visible:

- London 'Multicultural' is more typical than the multicultural classification in other parts of GB, probably due to its dominance in defining the class.
- 'Countryside' OAs tends to be typical of their super-group.
- 'Constrained by Circumstances' in Glasgow is less typical than in Newcastle.
- 'Typical Traits' and 'Prospering Suburbs' show low spatial autocorrelation, but the latter are restricted to Northern England and the Southwest.

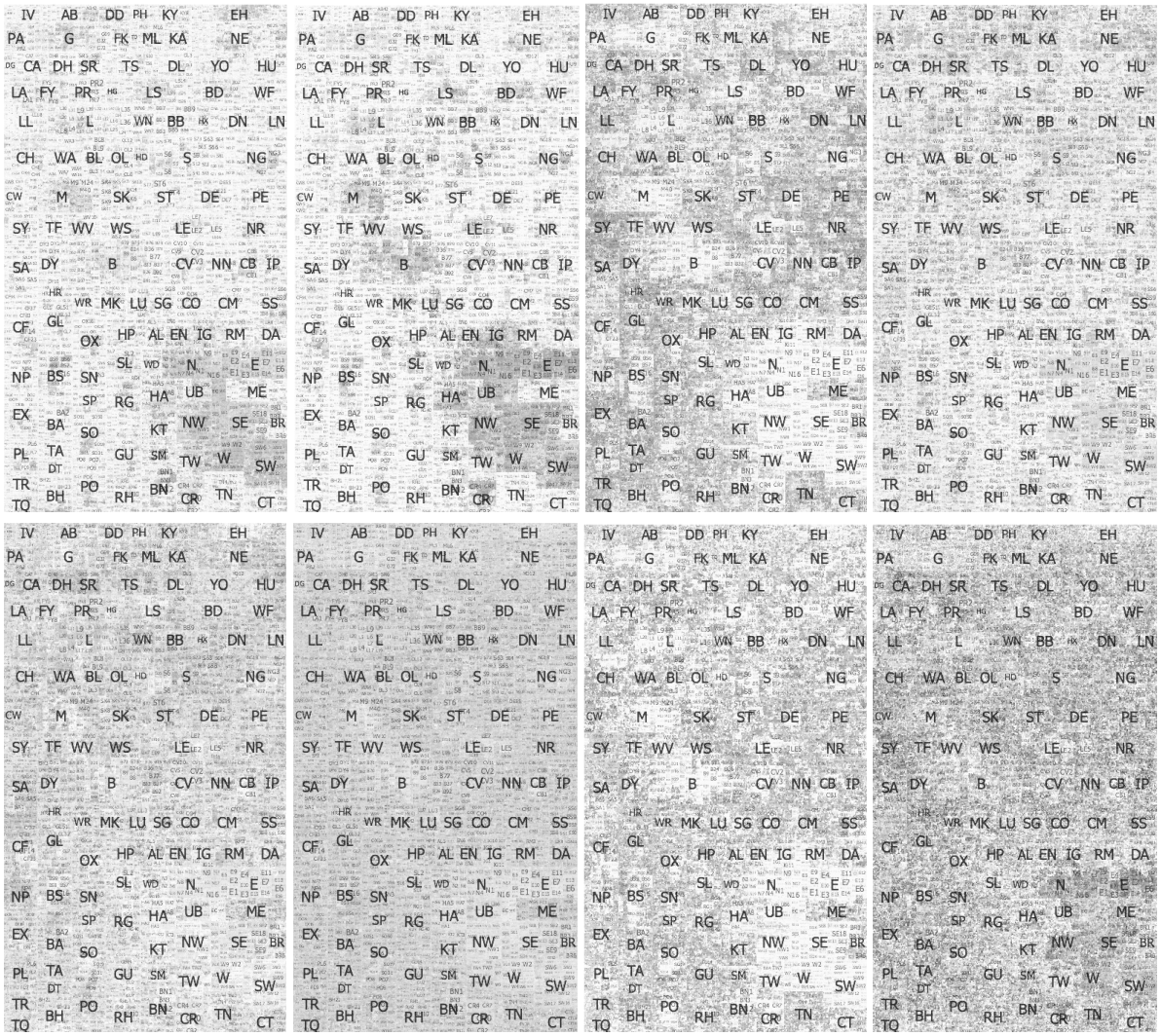


**Figure 4.** Similarity to specific super-groups (bottom right is similarity to closest, as Figure 3 but using lightness only to show typicality).

Mapping the similarity to each super-group individually is also interesting. Since lightness is comparable between the hues, direct comparison can be made:

- 'Multicultural': correlates to large cities with those in London most typical.
- 'Typical Traits': generally areas show a very strong similarity to this cluster, but less so in large cities.

- 'Countryside': tends towards either strong or weak similarity.
- 'Constrained by Circumstance': A broad-scale northern emphasis in part a response to the southern emphasis of some of the other classes

## 5. Conclusion and ongoing work

Population-density normalising cartograms with well-designed colour schemes that allow the representation of two variables, appears to be a useful combination of techniques. The space-efficiency provided by rectangular cartograms enlarges OAs of interest at the expense of position and contiguity. Contiguity is perhaps the most serious concern and we are working on new cartogram layouts that minimise contiguity distortion. Broad spatial patterns are detectable and the postcode hierarchy provides a useful spatial reference. Where lightness and saturation are held constant, the colours are less distinguishable from each other, but with the benefit that no colour is given undue prominence. Lightness can be varied to map to another variable and these lightnesses become comparable across hues.

## 6. Acknowledgements

The authors are grateful to the Willis Research Network for funding Aidan's position and for Dan Vickers' help with the assembling the OAC data. National Statistics Postcode Directory, obtained through UKBorders/Edina. OAC-data made available from SASI (2009).

## References

Brewer, C., 1994. Colour use guidelines for mapping and visualisation. In *Visualization in Modern Cartography*. New York: Elsevier, pp. 123-147.

DCLG (UK Department of Communities and Local Governement) (2009). Supporting local information and research: Understanding demand and improving capacity. *http://www.communities.gov.uk/publications/communities/supportinglocalresearch*

Gastner, M.T. & Newman, M.E.J. (2004). Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of USA*, 101(20), 7499-7504.

Florisson, S., Kreveld, M.V. & Speckmann, B. (2005). Rectangular cartograms: construction & animation. In *Proceedings of the 21st annual symposium on Computational geometry*. Pisa, Italy: ACM, pp. 372-373. Available at: http://portal.acm.org/citation.cfm?id=1064152

LCC (Leicestershire County Council) (2008). Report: Leicestershire Community Safety Partnership Strategic Assessment 2008. *http://www.lsr-online.org/reports/Leicestershire_community_safety_partnership_strategic_assessment_2008*

SASI (Social and Spatial Inequalities Group, University of Sheffield) (2009). The National Classification of Census Output Areas. *http://www.sasi.group.shef.ac.uk/area_classification/*

Slingsby, A., Dykes, J. & Wood, J. (2009). Configuring Hierarchical Questions to Address Research Questions. *IEEE Transactions on Visualization and Computer Graphics*, 15(6).

Vickers, D. & Rees, P. (2007). Introducing the National Classification of Census Output Areas. *Population Trends*, 125, 380-403.

Wijffelaars, M. Vliegen, R and van Wijk, J. (2008). Generating Color Palettes using Intuitive Parameters. *Computer Graphics Forum*, 27(3), 743-750.

Wood, J. & Dykes, J. (2008). Spatially Ordered Treemaps. Visualization and Computer Graphics, *IEEE Transactions on Visualization and Computer Graphics*, 14(6), 1348-1355.

## Biographies

Dr Jo Wood is a reader in geographic information at the giCentre at City University London with research interests in geovisualization, terrain modelling and object oriented programming for spatial sciences.

Dr Aidan Slingsby is a Willis Research Fellow at the giCentre at City University London with research interests in designing, implementing and using geovisualisation techniques for assessing data quality and variability and for visual data analysis.

Dr. Jason Dykes is a Senior Lecturer at the giCentre, City University London undertaking applied and theoretical research in, around and between information visualization, interactive analytical cartography and human-centred design.

# Surnames as Indicators of Cultural and Linguistic Regions in Europe.

James Cheshire[1], Pablo Mateos[1], Paul A. Longley[1]

[1]Department of Geography and Centre for Advanced Spatial Analysis, University College London.

james.cheshire@ucl.ac.uk, spatialanalysis.co.uk

KEYWORDS: Surnames, Europe, Clustering, Geodemographics, Lasker Distance.

## 1. Introduction

The study of names is a many-sided enterprise with great and exciting intellectual potential (Zelinksy, 1997). This is especially true of European surnames where high linguistic and cultural diversity have produced a rich variety of surnames. Previously these have been subject to relatively little large-scale spatial analysis (see Colantonio et al. (2003) and Cheshire et al (2009) for full reviews). This study seeks to establish the degree to which the spatial distributions of European surnames form recognisable regions when compared to well-known broad linguistic and cultural areas. The quantity of surnames and geographic extent of the data is unprecedented in this field of research (Manni et al., 2005). The results provide an interesting classification of 16 European countries that can be utilised in future research as a basis for hypothesis generation and smaller scale studies.

This study takes as a given that surnames vary over space and that these variations are culturally determined (Zelinksy, 1997). The focus here will be methodological by outlining an inductive approach to discovering and representing the regionalities in surname distribution that may exist across Europe. This approach utilises proven methods for the meaningful aggregation, hierarchical clustering and subsequent mapping of millions of surname locations obtained from telephone directories and censuses.

## 2. Methods

Surnames are commonly mapped individually or in groups according to a shared characteristic, such as patronymic ‘s’ names in Wales. These maps are appropriate for specific studies into a particular name or group but are inadequate for large scale, generalized studies. Through their interest in surnames as an indication of the genetic relationships between groups of people, geneticists have devised the Coefficient of Isonymy that provides a method of aggregating the information contained within the spatial locations of millions of surnames (Lasker, 1977). The Coefficient of Isonymy establishes the extent to which the same name (isonymy) occurs between the populations of two or more spatial units. It can be defined as:

$$\mathrm{Ri} = \sum_{i} \frac{\mathrm{piA\ piB}}{2} \tag{1}$$

where $p_{iA}$ is the relative frequency of the $i^{th}$ surname in population A and $p_{iB\,i}$ is the relative frequency of the $i^{th}$ surname in population B. A lack of similarity between very diverse populations will produce very small Coefficient of Isonymy values that are hard to interpret and handle computationally. The Lasker Distance (Rodriguez-Larralde et al. 1994) is an extension of the Coefficient of Isonymy that produces more useful results. It is simply defined as:

$$L_{iAB} = -\ln(2R_{iAB}) \tag{2}$$

The Lasker Distance values between spatial units can be thought of as distance in "surname space". Larger values between groups suggest greater difference and smaller values greater similarity in terms of surname composition. The results of the calculation can be treated as a dissimilarity matrix which provides a convenient input for the Ward's hierarchical clustering and multidimensional scaling (MDS). Space does not permit a full justification and in depth explanation of these methods. An in depth analysis of a variety of clustering methodologies in this context can be found in Cheshire et al. (2009).

Ward's (1963) grouping algorithm is a popular method of hierarchical agglomeration. The procedure forms hierarchical groups of mutually exclusive subsets in attribute space, each of which contains members of maximal similarity in terms of the specified characteristics (Ward, 1963). The algorithm begins by assigning the $n$ initial number of observations to $(n - 1)$ exclusive sets by considering the union of all possible $[n(n - 1)/2]$ pairs for the functional relation that matches an objective function chosen by the investigator, and then proceeds by successive iteration (Ward, 1963). As with other hierarchical classifications (see Gordon, 1987), the outcome of clustering can be visualised as a dendrogram that illustrates the relationship between each observation and the rest, where all of the observations are joined together at the "trunk of the tree". Each time two observations are joined, a new node is introduced with branches to the joined observations, the length of which are known as the cophenetic distance. This indicates the strength of the relationship between the observations (Kleiweg et al., 2004). Joining the clustering outcome to the boundary data enables the allocations to be shown as a choropleth map. Inspection of the resulting dendrogram with a view to allocate a number of clusters as close to the number of input countries (16) informed the decision to map 18 clusters.

In addition to hierarchical clustering, MDS was used to provide an effective summary of the degree to which surnames registered in the same country are clustered in multidimensional space. MDS is a well established method of reducing the dimensionality of a data set from an $m$ x $n$ matrix with a large value of $n$ to a similarity matrix with very few values of $n$ (Everitt et al., 2001). MDS is well suited to studies where the distance measures (in this case Lasker Distance between areas) arise directly from prior analysis (Everitt et al., 2001). In this study reducing n to 3 provided the maximal data reduction whilst minimising the loss of information. For ease of plotting a reduction of $n$ to 2 is shown. The results from the former are shown on the conference poster.

## 3. Data

The data used in this study are a subset from the database created for UCL's World Names Profiler (www.publicprofiler.org/worldnames) that contains the surnames and approximate locations of approximately 300 million people from 26 countries. Analyzed here are the 16 European countries, between them containing approximately 5,950,000 million unique surnames. Two levels of geography (NUTS 1 and NUTS 2) are used in this study. A list of the 16 countries and the Nomenclature of Territorial Units for Statistics (NUTS) level of geography used in the analysis is provided in Table 1. In the UK, for example NUTS 1 corresponds to the Government Office Regions (GOR), whilst counties correspond to NUTS 2. The variation in NUTS levels used by this research is due to a lack of data at NUTS 2 level for Serbia, Macedonia, the Netherlands and Norway. The use of NUTS 1 for the remaining countries was prompted by very low populations within many of the NUTS 2 spatial units for these areas. In total the Lasker Distances between 763 spatial units were calculated. All the surname and location data were derived from publicly available telephone directories or national electoral registers from the 2000-2005 period. To our knowledge, no study of this kind has been completed on a continental scale before with this quantity of unique surnames.

## 4. Results and Discussion

The mean Lasker Distance between the 763 spatial units was 10.46, with a range of values between 1.66 and 19.68. The smallest distances often occurred between contiguous areas and the largest distances across international boundaries.

The exclusion of spatial information (such as contiguity constraints, or distance weightings) from the hierarchical clustering makes the spatial uniformity of the clusters shown in Figure 1 especially impressive. Unlike Belgium which has been divided into a northern area of Dutch surnames and a southern area of French surnames, the multilingual countries of Luxembourg and Switzerland have been given unique cluster allocations rather than being partitioned along their known linguistic divisions. This suggests that a greater degree of surname mixing has occurred between the different linguistic areas of these countries than occurs between Wales, Scotland and England where each have been assigned a separate cluster allocation. Although not present in the 18 cluster outcomes, transitions in surname compositions are present along other linguistic boundaries, such as between Catalan areas of Spain and the rest of the country, but these require the dendrogram to be partitioned into a greater number of clusters,

**Table 1:** A list of the countries, and their respective level of spatial granularity.

| Country | NUTS Level |
|---|---|
| Denmark | 1 |
| Netherlands | 1 |
| Poland | 1 |
| Serbia and Macedonia | 1 |
| Sweden | 1 |
| Austria | 2 |
| Belgium | 2 |
| France | 2 |
| Germany | 2 |
| Ireland | 2 |
| Italy | 2 |
| Luxembourg | 2 |
| Norway | 2 |
| Spain | 2 |
| Switzerland | 2 |
| UK | 2 |

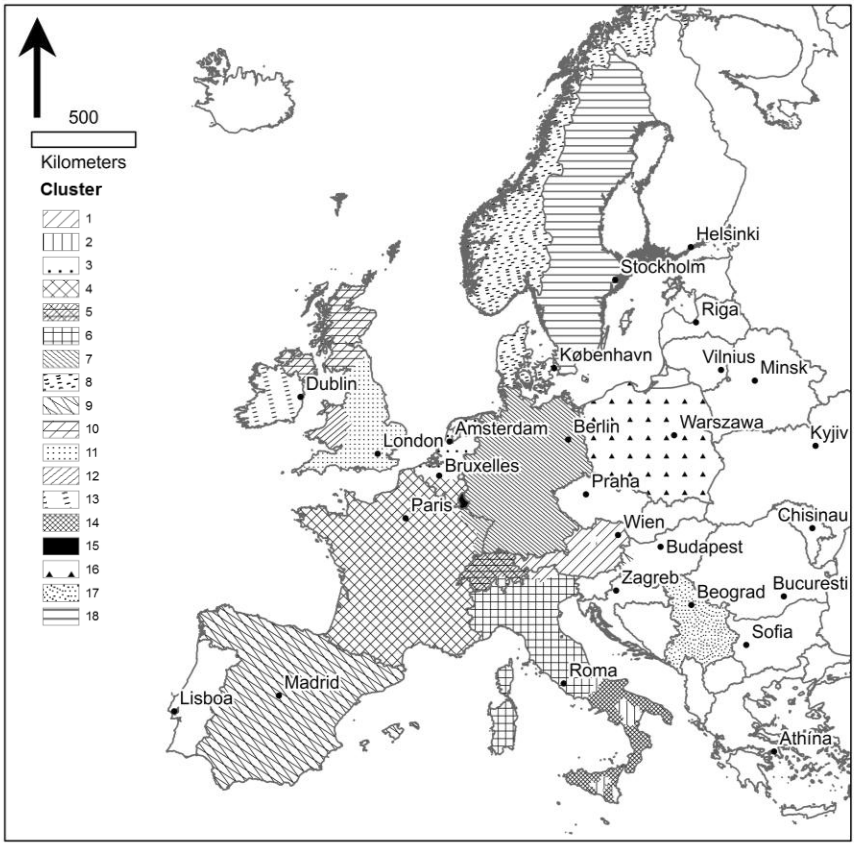



**Figure 1:** A map of the 18 cluster allocations produced from the Ward's Hierarchical Clustering of Lasker Distances. Each allocation is represented as a unique pattern. Cophenetic distances between adjacent clusters can be large, as is the case between Poland and Germany, or relatively small such as between England and Wales. Areas of no-data are white.

suggesting more subtle differences in surname composition between these areas. Of the Scandinavian countries, Norway and Denmark are more similar to each other than to Sweden. Italy's surname composition appears fragmented with a north/ south split. The former includes Rome and Sardinia, whereas the later has been split into two groups with the province Basilicata and part of Sicily forming one cluster allocation and the rest of Scilly and Southern Italy forming the second.

The MDS plots in Figure 2 provide an effective means of gauging the similarity between spatial units within a country. From these it is clear that some countries have tighter distributions than others and these can be characterised as having a unique, in the context of Europe, but uniform composition of surnames within their borders. Ireland, Poland, Norway and Germany appear the most tightly clustered. By contrast countries where multiple languages are spoken, such as Switzerland, Luxembourg and Belgium, have more dispersed MDS distributions. In France, outliers include the more Germanic areas surrounding Alsace. Finally, although Serbia and Montenegro have been allocated a single cluster in Figure 2 their points in multidimensional space appear relatively distant from each other. This may be a product of the large spatial units and the high relative difference between the three areas of Serbia and Montenegro and the rest of Europe.

It is recognised that naming conventions vary across Europe. It was felt unnecessary to account for this as the purpose here is to simply identify areas of similarity/ difference in surname compositions. For greater meaning to be attached to these results, such as genetic relatedness, conventions need to be accounted for in the initial Lasker Distance calculation. The unsurprising nature of many of the surname regions highlighted (judged by their conformity to well-known national and linguistic boundaries) provides strong evidence that the inductive approach of this study, as demonstrated through its data and methods, is appropriate when attempting to establish the existence of regional patterns in Europe's surname distributions. A great deal more variation exists beyond the 18 groupings outlined above which will be the subject of future research. In addition, the data quality, spatial scale and extent can be improved through additional cleaning of the database, geocoding the available address data to NUTS 3 or finer levels of granularity and obtaining data for the countries where it is lacking. The sensitivity of the analysis to the population sizes of the input spatial units also requires further investigation. This appears to be particularly important on an international level where the spatial units and their population sizes vary. Finally, many interesting patterns and distributions emerge on a national level that are beyond the scope of this paper, but could be easily investigated with the data and methods demonstrated by this research**.**

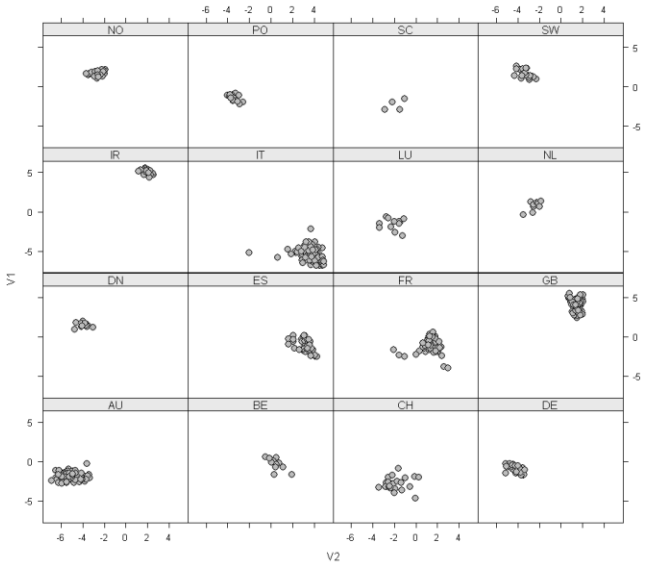



**Figure 2:** Plots produced from the 2-dimensional MDS for each of the 16 countries. From top left the countries are: Norway (NO), Poland (PO), Serbia and Macedonia (SC), Sweden (SW), Ireland (IR), Italy (IT), Luxembourg (LU), Netherlands (NL), Denmark (DN), Spain (ES), France (FR), United Kingdom (GB), Austria (AU), Belgium (BE), Switzerland (CH), Germany (DE).

## 5. Acknowledgements

The authors would like to thank the work of Muhammad Adnan and Maurizio Gibin for their work in assembling the database and linking it to the boundary data. This project was undertaken as part of James Cheshire's ESRC CASE PhD Studentship in Collaboration with ESRI (UK). The reviewer's comments were gratefully received and we hope have been fully addressed in the improvements made to the original paper and its associated poster.

## 6. References

Cheshire, J., Mateos, P. Longley, P. 2009. Family Names as Indicators of Britain's Regional Geography. *CASA Working Paper 149.* Available from http://www.casa.ucl.ac.uk/publications/workingpapers.asp.

Colantonio, S., Lasker, G., Kaplan, B., Fuster, V. 2003. Use of Surname models in Human Population Biology: A Review of Recent Developments. *Human Biology.* 75, 6: 785-787.

Everitt, B., Landau, S., Leese, M. 2001. *Cluster Analysis 4th Edition.* Hodder, London.

Gordon, A. 1987. A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. Series A (General).* 150, 2: 119-137.

Kleiweg, P., Nerbonne, J., Bosveld, L. 2004. Geographic Projection of Cluster Composites. In Blackwell, A., Marriott, K., Shimojima, A. *Diagrams 2004, Lecture Notes in Computer Science.* Springer, New York.

Lasker, G. 1977. A Coefficient of Relationship By Isonymy: A Method for Estimating the Genetic Relationship Between Populations. *Human Biology.* 49, 3: 489-493.

Manni, F., Toupance, B., Sabbagh, A., Heyer, E. 2005. New Method for Surname Studies of Ancient Patrilineal Population Structures, and Possible Application to Improvement of Y-Chromosome Sampling. *American Journal of Physical Anthropology.* 126: 214-228.

Rodriguez-Larralde, A., Pavesi, A., Siri, G., Barrai., I. 1994. Isonymy and the Genetic Structure of Sicily. *Journal of Biosocial Science.* 26: 9-24.

Ward, J. 1963. "Hierachical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* 58, 301:236-244

Zelinsky, W. 1997. Along the Frontiers of Name Geography. *Professional Geographer.* 49, 4: 465-466.

## 7. Biographies

***James Cheshire*** *is halfway through his ESRC CASE PhD studentship (in collaboration with ESRI (UK)) in UCL's Department of Geography. His research focus is the spatial analysis of surnames and its applications. He is a also a research assistant on the Wellcome Trust's "People of the British Isles" project. James' research can be followed at spatialanalysis.co.uk.*

***Paul Longley*** *holds a chair in Geographic Information Science at UCL and acts as Deputy Director of CASA. His publications include twelve books and more than 125 refereed journal articles and contributions to edited collections. He is a co-I on a node of the ESRC National Centre for E Social Science and a co-editor of the journal Environment and Planning Series B.*

***Pablo Mateos*** *is Lecturer in Human Geography in the Department of Geography at University College London (UCL). His research interests lie within Population and Urban Geography and his work focuses on investigating ethnicity, migration and socio-spatial inequalities in contemporary cities. PhD in Geography (UCL 2007); MSc in GIS and Human Geography (University of Leicester 2004).*

# The application of geodemographics to social vulnerability and volcanic hazard assessment

Iain Willis[1] Maurizio Gibin[1] Joana Barros[1] Richard Webber[2]

[1]*Department of Geography, Environment and Development Studies, Birkbeck College, Malet Street,* London, WC1E 7HX
Tel: 020 7631 6473 | Fax: +44 (0) 20 7631 6498
Email iwilli02@bbk.ac.uk, m.gibin@geography.bbk.ac.uk, j.barros@bbk.ac.uk

[2]*Department of Geography, Kings College London, Strand*, London, WC2R 2LS
Tel: +44 (0)20 7836 5454 | Fax: +44 (0) 20 7631 6498
Email : richardwebber@blueyonder.co.uk

## 1.0 Introduction

The dynamic forces of urbanisation that characterised much of the 20th Century and still dominate population growth in developing countries have led to the increasing risk of natural hazards in countless cities around the world (Chester 2000, Pelling 2003). None of these physical dangers is more tangible than the threat volcanoes pose to the large populations living in close proximity. Vesuvius, a recognised decade volcano (IDNDR 1990) has an estimated 550,000 people that live in areas at risk from Pyroclastic Density Currents (PDC) (Barberi 2008) and a further 4 million living in the vast suburbs of neighbouring Naples. Though quiescent since 1944, Vesuvius remains the only volcano to have erupted on mainland Europe in the last 100 years.

Coupled with the geophysical risks of the Somma Vesuvius region, a 38% population growth in Italy from 1901 to 2001 mean an eruption of Vesuvius is now a far more devastating proposition than ever before.

### *1.1 Evacuation*

Current evacuation plans divide the area around the volcano into three zones that are designated for different levels of priority and risk; Yellow, Blue and Red zones delineate the evacuation regions and were designed by the Department of Civil Protection (DPC 1995). However, the evacuation plans have been criticised in recent years as lacking local understanding, support and confidence (Barberi 2008). These plans have not been updated since 1995.

The most hazardous region, classified as the Red zone, contains all the municipalities around the volcanic vent and is home to 550,000 residents. This area of Italy has one of the highest population densities in Europe (Eurostat 2010).

### *1.2 Social Vulnerability*

Unemployment has fallen recently in the Campania region but compared to Italian national averages, this region is still one of the most deprived areas in Italy (European Commission 2009). Therefore, the consequences of an eruption have socio-economic consequences that go far beyond just the physical risks associated with volcanoes.

This study draws from the literature on natural disaster risk (Quarantelli 1978, Hewitt 1983, Wisner 2001), whereby demographic and socio-economic variables can be recognised as increasing an individual's social vulnerability during a disaster. The use of demographic data to assess social vulnerability came about during the late 1970's, early 1980's as a paradigm

shift in the standard interpretation of natural disasters (Wisner 2004); the classical view regarding 'natural disasters', such as earthquakes, volcanoes and hurricanes as the consequence of natural processes. However, following the work of early pioneers in disaster management, focus began to be placed on understanding how hazards became 'disasters'.
The fundamental links and inter-dependencies between social marginalisation, an individual's access to resources, and their capacity to financially recover from a disaster are all themes explored here as a new methodology is proposed for social vulnerability assessment.

Based on Cutter's work (2000) creating a Social Vulnerability Index (SoVi) using social statistics, this study uses geodemographic data for Italy to rank census regions using the same underlying princples of DRR.

During the onset of a natural disaster, certain socio-economic and demographic factors can affect an individuals propensity to risk.

For example, the elderly and the young are more vulnerable during disaster evacuation (McMaster 1988, O'Brian and Mileti 1992) and ethnic minorities may have less access to the same resource and political power (Pulido 2000) as indigenous populations.
Less affluent households are more likely to struggle in terms of their financial resilience and subsequent economic recovery following the onset of a disaster (Burton et al 1992).
Population density is another important consideration during evacuation measures (Johnson and Zeigler 1986) as more crowded areas are more difficult to evacuate.

Based on the premise that social statistics help understand a regions vulnerable population groups during a disaster, this paper addresses the possibility that geodemographic data can be used to show these same vulnerabilities at a micro-scale.

## 2.0 Creating the SoVi

Experian kindly loan the use of MOSAIC Italy 2007 data for this study. The data has then been combined with geophysical risk maps to create a SoVi for the area around the volcano. For the purposes of spatial modelling, a large Sub-Plinian eruption is hypothesised for Mt Vesuvius.

Drawing from the literature on disaster risk reduction (DRR), each of the 223 MOSAIC variables provided is assessed according to its ability to discriminate a household's social vulnerability to evacuation, access to resource, financial recovery and physical risk of house collapse. A range of social statistical methods were used to assess each of these variables. This includes gini-coefficients, Pearson's correlation coefficients and the index range of survey variables (Leventhal 1995). These factors are then weighted and combined with geophysical risk modelling of a Sub-Plinian eruption to formulate a SoVi for Vesuvius. It should be noted that the weighting methodology used in this assessment has never been used before and is proposed as a means of appropriately weighting geodemographic variables in the SoVi.

| Social vulnerability factors | MOSAIC Italy Variables |
|---|---|
| Individual/Household access to resources during a disaster | Accessibility to local facilities, Phone connection, Ethnicity, Rurality |
| Demographic/Ethnic vulnerability during evacuation | Gender, Ethnicity, Age, Daily movements, |
| Financial capacity to recover | Socio-economic (income, rented, loan, credit cards) |

| Category of social vulnerability, Mt Vesuvius area | MOSAIC Italy Variables |
|---|---|
| House/Flat is vulnerable to collapse (pyroclastic flow / tephra) | Building age, Building type, High rise flats |

**Table 1:** Link between social vulnerability and MOSAIC data

*2.1 Measurements of discrimination*

The degree to which geodemographic classifications can discriminate between MOSAIC profiles is a key factor in their ability to describe neighbourhoods. Therefore, the following statistical tests allowed for a comparative study.

*Index Range :*

Index Range = Index Value $_{\text{Max x}}$ – Index Value $_{\text{Min x}}$

*Index range (Leventhal 1995)*

*Lorenz Curves*

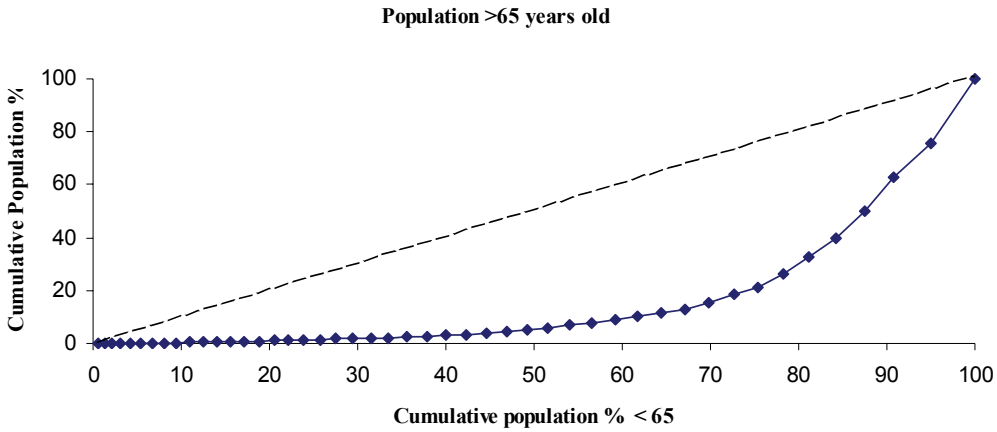



**Figure 1**: Lorenz curve, MOSAIC Italy variable: Population >65 years old (Mosaic Italy 2007)

Gini coefficients were calculated for each of the 24 variables. These values can only range from 0-1 and are independent of whether the final value is positive/negative. If the coefficient is closer to 0 than 1, the more evenly distributed the variable. If the coefficient is closer to 1, there is a more unequal distribution of a variable.

Results indicated the most unequal distribution of data included the following MOSAIC variables; % *Divorce, % Buildings with 3-10 flats* and *% Houses without water/toilet*.

*2.2 Correlations (Evacuation, Financial recovery, Access to resources)*

To assess the inter-dependencies of MOSAIC variables and reduce data redundancy within a risk category, Pearson correlations were calculated. Correlation coefficients vary between -1 and 1. The closer a value is to 1 or -1, the greater the linear correlation between variables.

| Variables | % Separated | % Widowed | % Divorced | % Age <5 | % Age >65 | % Daily movement (inside Comune) | % Buildings with 3-10 flats | % Buildings with more than 10 flats | People per Household | Household Density | Population Density |
|---|---|---|---|---|---|---|---|---|---|---|---|
| % Separated | | 0.260 | **0.967** | -0.170 | 0.180 | 0.559 | 0.418 | 0.616 | -0.670 | 0.579 | 0.466 |
| % Widowed | 0.260 | | 0.375 | -0.669 | **0.970** | -0.360 | -0.070 | 0.040 | -0.752 | 0.200 | 0.100 |
| % Divorced | **0.967** | 0.375 | | -0.280 | 0.318 | 0.526 | 0.367 | 0.585 | -0.753 | 0.568 | 0.435 |
| % Age <5 | -0.170 | -0.669 | -0.280 | | -0.682 | 0.210 | 0.230 | -0.130 | 0.485 | -0.130 | -0.060 |
| % Age >65 | 0.180 | **0.970** | 0.318 | -0.682 | | -0.346 | -0.140 | 0.010 | -0.745 | 0.160 | 0.050 |
| % Daily movement (inside Comune) | 0.559 | -0.360 | 0.526 | 0.210 | -0.346 | | 0.407 | 0.715 | -0.030 | 0.601 | 0.593 |
| % Buildings with 3-10 flats | 0.418 | -0.070 | 0.367 | 0.230 | -0.140 | 0.407 | | 0.130 | -0.130 | 0.210 | 0.230 |
| % Buildings with more than 10 flats | 0.616 | 0.040 | 0.585 | -0.130 | 0.010 | 0.715 | 0.130 | | -0.220 | **0.847** | 0.844 |
| People per Household | -0.670 | -0.752 | -0.753 | 0.485 | -0.745 | -0.030 | -0.130 | -0.220 | | -0.356 | -0.190 |
| Household Density | 0.579 | 0.200 | 0.568 | -0.130 | 0.160 | 0.601 | 0.210 | **0.847** | -0.356 | | **0.978** |
| Population Density | 0.466 | 0.100 | 0.435 | -0.060 | 0.050 | 0.593 | 0.230 | 0.844 | -0.190 | **0.978** | |

**Table 2:** Evacuation variable $R^2$ correlations

*Table 2* shows the correlations between variables that would increase a household's vulnerability during disaster evacuation. Variables with a correlation of close to 1 were taken out of the SoVi because their correlation was too great. This would have resulted in data redundancy and effectively over representing a variable.

*2.3 A new methodology for SoVi*

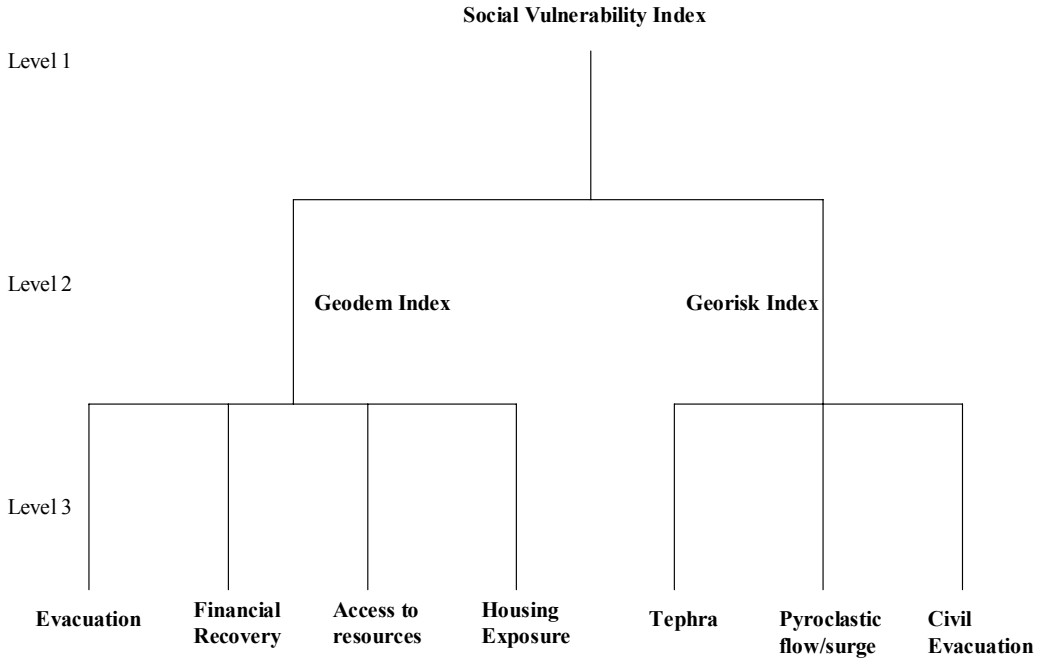



**Figure 2:** SoVi hierarchy

Level 3
These are the individual social vulnerability scores for each household for the following social and physical risks; *Evacuation, Financial recovery, Access to resources, Building exposure, Tephra fallout, Pyroclastic surges and Civil Evacuation (DPC 1995).*

Level 2
These index scores were created as a composite of the respective social and physical risk scores.

Level 1
The overall SoVi is calculated as an index from all physical and social variables in level 1.

To factor in a level of discriminatory weighting in this methodology, Gini-coefficients were used for each MOSAIC variable. This meant variables with Gini coefficients closer to 0 were given less weighting in the overall vulnerability score.

The following four equations describe how the index scores were calculated as a metric for each variable ($_x$). These were then combined for all social factors to create the *Geodem Index*.

1. Weighted variable $_{x}$ = ([1]Gini-coefficient $_{x}$)* (MOSAIC Index Value $_{x}$)
2. $\sum$ (Weighted variable $_{xn}$) = Vulnerability score $_{x\ (Area\ of\ social\ vulnerability)}$
3. Total Social Vulnerability$_x$ =$\sum$ (Vulnerability scores$_x$) ($_{Evacuation,\ Financial\ recovery,\ Access\ to\ resources,\ Building\ exposure}$)
4. Geodem Index = Total Social Vulnerability$_x$ / Total Social Vulnerability $_{MeanAverage}$ * 100

In order to calculate the overall SoVi, it was necessary to first calculate risk ranks for areas subject to Tephra, Pyroclastic surges and the Civil Evacuation around the volcano. For simplicity with regards to the index model, a numeric risk number between 0-3 was assigned for each Census area and for each hazard (3=high risk, 0=Very low/no risk). These value were then multiplied by a factor of 10 and accumulated to provide a *Georisk Index*.

The final SoVi score was deduced from all social and physical scores as an index.

Social Vulnerability Index$_x$ =$\sum$ (Social & Physical vulnerability scores$_x$) ($_{Evacuation,\ Financial\ recovery,\ Access\ to\ resources,\ Building\ exposure,\ Tephra,\ Pyroclastic\ surge,\ Civil\ Evacuation\ areas}$)

## 3.0 Results

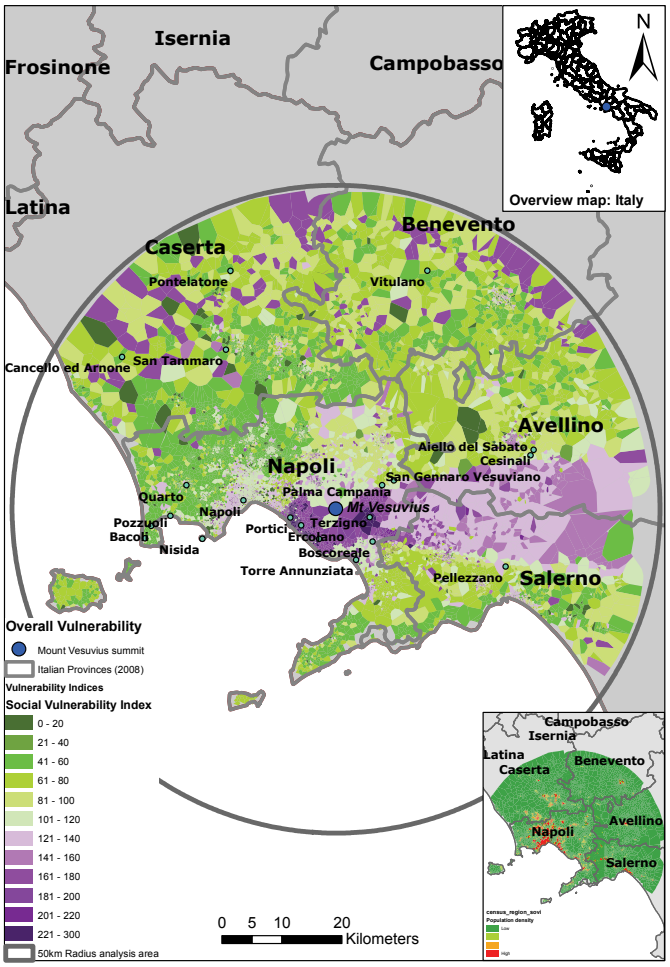



**Figure 3**: SoVi Mt Vesuvius (50km analysis zone)

## 4.0 Project summary and considerations

Coastal areas in the Napoli province are some of the most densely populated in Europe. They are also the areas identified from this research as those impacted the hardest from an eruption of Vesuvius. This study has identified particularly vulnerable demographic groups within this area. *Low-income*, *urban areas* with *eldery* populations are the most vulnerable geodemographic categories to the social consequences of a volcanic eruption.

It is believed this work holds value in contributing towards both understanding the spatial orientation of social vulnerability while proposing a transferable methodology for other natural hazards.

This model serves as a practical and micro-level assessment to social risk. Civil protection agencies and NGO's could make use of a lower level of granularity for disaster mitigation and management. Community outreach and hazard awareness could be targeted more carefully to those demographic profiles worst affected. Local authority planning and development projects could use these maps to mitigate the spatial variance of risk around a natural hazard. With current Civil Protection plans for Vesuvius considered by 60% of residents in the most hazardous area of the province to be inadequate (Barberi 2008), this work could contribute to the necessary reassessment of evacuation measures.

The use of geodemographics for DRR should be treated with care to fully understand the limitations and caveats of using classification system data as misinformation can be more harmful than no information.

The MOSAIC Italy 2007 data used for this work is nearly 3 years old and data accuracy has therefore degraded in the years following its original release. Many demographic clusters defined in this dataset may not be entirely valid or representational of contemporary Neopolitan populations.

Likewise, data associated with MOSAIC Italy is derived from telemarketing surveys, which are unlikely to completely reflect the rich profile of all residents in the Campania region. Geodemographics are also subject to the rigours of aggregated data (MAUP), and therefore, many census region characteristics may not accurately portray their residents.

However, though geodemographics do not provide a panacea solution to the complexity of DRR modelling, a new micro-scale of assessment for social vulnerability would be a welcome addition to disaster preparedness, mitigation and overall reduction.

Acknowledgements
This work has been carried out with the support of Birkbeck, University of London. Many thanks also go to Experian for the use of MOSAIC Italy 2007, without which this study would not have been possible.

**References**

Barberi, F., Davis, M.S., Isaia, R., Nave, R., Ricci, T. (2008). "Volcanic risk perception in the Vesuvius population." Journal of Volcanology and Geothermal Research 172: 244–258.

Burton, I., Kates, R.W. and White, G.F. 1992: The environment as hazard. New York: Oxford University Press.

Cioni, R., Bertagnini ,A., Santacroce, R., Andronico, D. (2008). "Explosive activity and eruption scenarios at Somma-Vesuvius (Italy): Towards a new classification scheme." Journal of Volcanology and Geothermal Research 178: 331-346.

Cutter, S. L., J. T. Mitchell, et al. (2000). "Revealing the Vulnerability of People and Places: A Case Study of Georgetown County, South Carolina." Annals of American Geographers 90(4): 713-737.

Chester, D.K., Degg, M., Duncan, A.M., Guest, J.E. (2000) "The increasing exposure of cities to the e!ects of volcanic eruptions: a global survey" Global Environmental Change B: Environmental Hazards 2 (89-103) 2000 - Elsevier

DPC, 1995. Pianificazione Nazionale d'Emergenza dell'Area Vesuviana. Dipartimento della Protezione Civile, Roma.

Esposti Ongaro, T., Neri, A., Menconi, G., de'Michieli Vitturi, M., Marianelli, .P, Cavazzoni, C., and G. Erbacci, Baxter, P.J. (2008). "Transient 3D numerical simulations of column collapse and pyroclastic density current scenarios at Vesuvius." Journal of Volcanology and Geothermal Research 178: 378–396.

Eurostat, European Commission, 2009. Available: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ (Accessed 15 November 2009)

Experian Mosaic (2009) *Index.* Available at: http://strategies.experian.co.uk/. (Accessed 15 November 2009)

ISTAT. (2001). *Istituto Nazionale di Statistica.* Available: http://www.istat.it/english/. (Accessed 09/06/2009)

Hewitt, K. (1983). "Interpretations of calamity from the viewpoint of human ecology." Winchester M.A. Allen & Unwin

Johnson, J., and Zeigler, D. 1986. Evacuation Planning for Technical Hazards: An Emerging Imperative. Cities 3: 148-156.

Leventhal B (1995). "Evaluation of geodemographic classifications." Journal of targeting, measurement and analysis for marketing. 1995; Volume 4: 173-183.

McMaster, R.B. and Johnson, J.H., jr, 1887: Assessing community vulnerability to hazardous materials with a geographic information asystem. In Chrisman, N.R., editor, Autocarto 8 Proceedings, Falls Church, VA: American Society for Photogrammetry and American Congress on Surveying and Mapping, 471-80.

Pelling, M. (2003). Natural Disasters and Development in a Globalizing World: In a Globalizing World, Routledge, London.

Quarantelli, E.L (1978). " Disasters : theory and research". Sage Publications, c1978. 282 p, London ; Beverly Hills,

Pulido, L. 2000. "Rethinking Environmental Racism: White Privilege and Urban Development in Southern California." Annals of the Association of American Geographers 90:12–40.

Wisner, B. (2004). At risk : natural hazards, people's vulnerability, and disasters. London ; New York, Routledge.

**Biography**

*Joana Barros is a lecturer in GIS and MSc GI Science course Director at Birkbeck, University of London. Research interests include computational models of urban systems with a focus on the dynamics of urban growth in Latin American cities.*

*Maurizio Gibin works at both Birkbeck and UCL. Maurizio has previously been involved in developing geodemographic and social marketing systems to precisely target health campaigns in a cost effective way.*

*Richard Webber is a currently a visiting professor at Kings College, London. Richard developed the UK's first geodemographic classification system, MOSAIC. Current research interests have also included the analysis of people's ethnic origin based on surnames.*

*Iain Willis is a part-time GIS research student at Birkbeck, University of London. With a background in Physical Geography and GIS, current research interests are focused around the use of GIS to assess social vulnerability in hazardous areas.*

# "The Isolated State": an ABM approach to the von Thünen Model

S. Pottage[1], J. Barros [2], J. Nixon[3], A. Cannell[4]

Department of Geography, Environment and Development Studies
Birkbeck, University of London, Malet Street, London, WC1E 7HX
Telephone: +44(0)2070790644 | Fax: +44(0)2076316498
[1]Email: shaman.pottage@googlemail.com
[2]Email: j.barros@bbk.ac.uk
[3]Email: jez.nixon@yahoo.co.uk
[4]Email: alicannell@hotmail.com

KEYWORDS: Von Thünen, agent-based model, land cost, transportation, spatial constraints

## 1. Introduction

In 1826, Johann Heinrich von Thünen developed an agricultural model called "The Isolated State". This model illustrated the balance between land and transportation costs, assuming that agricultural land costs would decrease as the distance from the market (centre of town) would increase, and that the price of transportation would behave in the opposite way (von Thünen, 1826) (see **Figure 1**).

Despite being created in 1826 for agricultural land, the von Thünen model is still a very important model in geography. The model has served as the basis for urban land economics upon which a number of subsequent theories have been developed (see Alonso, 1964; Otter et al, 2001). A number of criticisms to this model have been made along the years, due to the model's homogeneity in the decision making process, agents, and land characteristics (Brown, 2006) and perfect market behaviour. Yet, the simplicity in which the model describes the land market behaviour by using transport and land cost as a function of the distance from the market made it one of the most important models in land economics to date.

In 1998, Philip Steadman created a "simple sketch" of the von Thünen model with a graphical user interface (GUI) for interaction (Steadman, 1998). His version of the model was created as a visualisation tool to teach students about land use economics (Steadman, 1998). The novelty of Steadman's model lies on the inclusion of modern forms of transport such as rail as well as natural constraints to development such as rivers and lakes.

In 2003, Sasaki and Box created an agent based version of von Thünen's model using the SWARM simulation system. They have used an agent-based model as a tool to relax the known restrictions of traditional market models (Parker and Filatova, 2008). Their focus was to verify von Thünen's theory in light of complexity theory concepts such as 'feedback', 'emergence' and 'lock in' (for more details see Sasaki and Box, 2003). Their model has successfully verified that von Thünen's concentric rings were the emergent result of a bottom-up process in which individual farmer's behaviour has collectively produced a spatial pattern.

While Steadman's model is an excellent visualisation tool and brings modern transport to von Thünen's theory, it does not address issues such homogeneity as Sasaki and Box's model does by recreating the model as an agent-based model. Thus, the present research aims to combine the contributions of both Steadman's and Box and Sasaki's work into a single model. This paper discusses the ongoing development of an agent-based version of von Thünen's Model, with the study of real world applications of von Thünen's theory as the ultimate aim.

This is to be achieved in a two step process. In the first stage, an agent based model (ABM) of the von Thünen theory was created using StarLogo software (MIT Media Laboratory, 2006) as a proof of

concept. StarLogo is a LOGO programmable modelling environment of decentralized systems which allows the user to examine and control the geospatial phenomenon (Resnick, 1998). In a second stage, the model will be recreated in either JAVA programming language using the programming platform RePast (Recursive Porous Agent Simulation Tootkit) (Argonne National Laboratory, 2007) or using the NetLogo Platform. This second version will introduce environmental and terrain conditions on the homogenous land and will concentrate on further development of a railway station that could eventually develop into satellite markets. In addition to further development from stage one, stage two will focus on the tests for sensitivity and validation of the model.

## 2. The von Thünen Model

Von Thünen developed an agricultural model called "The Isolated State", the primary assumptions of this model focused on the market being the centre of the town and the economic importance was based on the prediction of land use and land rent (von Thünen, 1826). Other assumptions of the model included; a) a wilderness outside the areas which can be cultivated, b) the land is flat and has no environmental features, and c) the soil quality and climate are consistent. In his theory, the farmers used only an oxcart to transport goods to the market across the land and behaved rationally to maximize profits. Von Thünen represented this mathematically using the equation below.

$$R = Y(p - c) - YTd \quad (1)$$

(Where R = Land Rent; Y = Yield per unit of land; c = Production Cost per unit of commodity; p = Market Price per unit of commodity; T = Transport Rate; d = Distance to market)

Four concentric rings developed surrounding the central market, each ring corresponding to a specific type of agriculture. Zone 1 (a) is land used for dairy products and intensive farming of perishable goods; Zone 2 (b) was utilised for timber and firewood; Zone 3 (c) was used for the agriculture of extensive field crops such as grains for bread, and finally the area outside zone 3, Zone 4, contained the rearing and breeding of animals, this is illustrated in the diagram in

**Figure 1**.

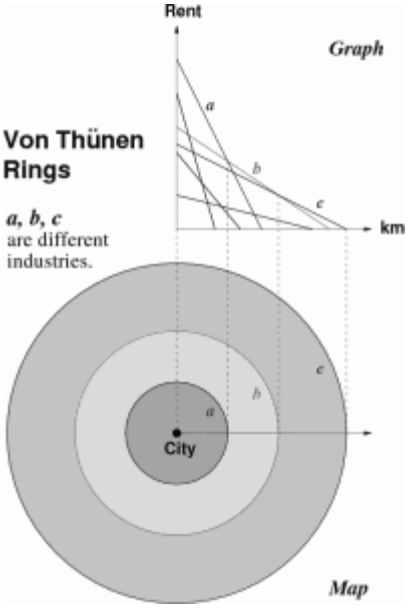



**Figure 1:** A simple pictorial description of the theory of land use as predicted by von Thünen.
(Source: http://www.prosper.org.au/wp-content/uploads/2008/08/thunen11-201x300.gif )

## 3. Model Development

Steadman's (1998) and Box and Sasaki's (2003) work has provided a good basis for the present model. The model focus is to look at the complexity of the farmers and their interaction with their environment and to be able to "overlay" an abstract map with natural features to visualise the extent at which this could be achieved. The von Thünen's model pre-dates modern innovations such as electrical refrigeration and transportation and these have had a dramatic effect on the pattern of agriculture and where farmers, in this instance, are located.

StarLogo (MIT Media Laboratory, 2006) is an agent based modelling system was used in this research to geovisualise the concentric rings of the von Thünen theory, to assess how the agents arrived to their end point and to gain an insight into the mechanisms for that change by varying the speed at which the agents interact with the environment.

In the model the agents each represent a desire or need to produce a particular commodity i.e. the produce which they will sell. The agents are a member of one of four breeds (a) perishables; (b) timber; (c) grain; and (d) ranching.

The land rent and a distance from the market attribute are included in each unit of land and at the start of the model all land rent equals zero, but as each agent tries to establish production for its commodity on that land unit, a bidding war ensues. The agent carries the knowledge of what yield, production cost, transport rate, and market price are associated with its' particular commodity. Then they are assigned, randomly, to cross the land bidding for patches and establishing farms before moving on to do the same elsewhere, this is shown as a flowchart in figure 2.

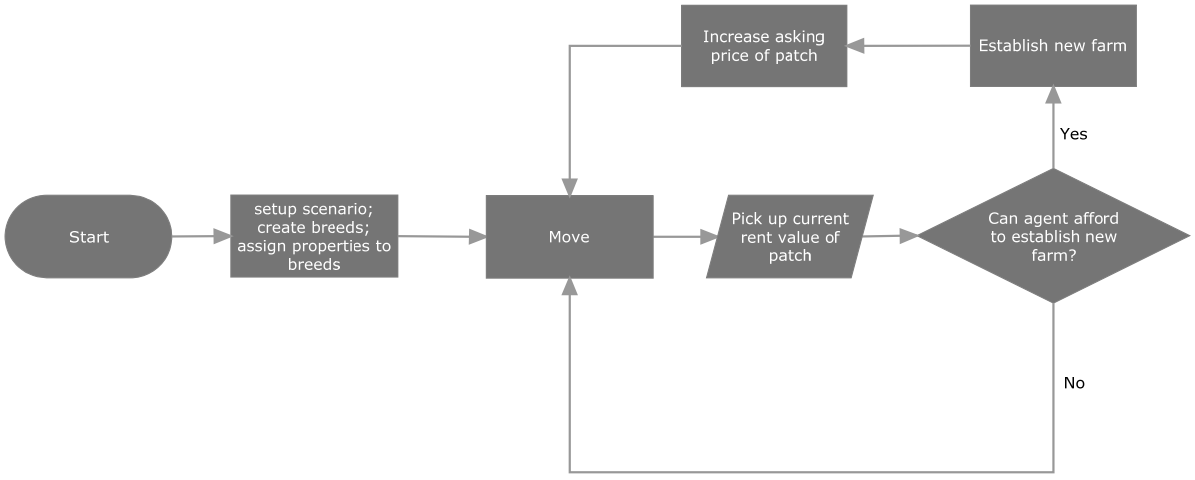



**Figure 2:** The Model Flowchart

The initial results of this model did not settle as predicted, as most competition for land would take place close to the market where all four commodities could be viably produced, so a random movement of the agents was included in the model and aptly named the "gravitational wiggle" as shown in figure 3. This feature prevented the agents from competing too close to the market and thus the concentric circles associated with von Thünen were formed.

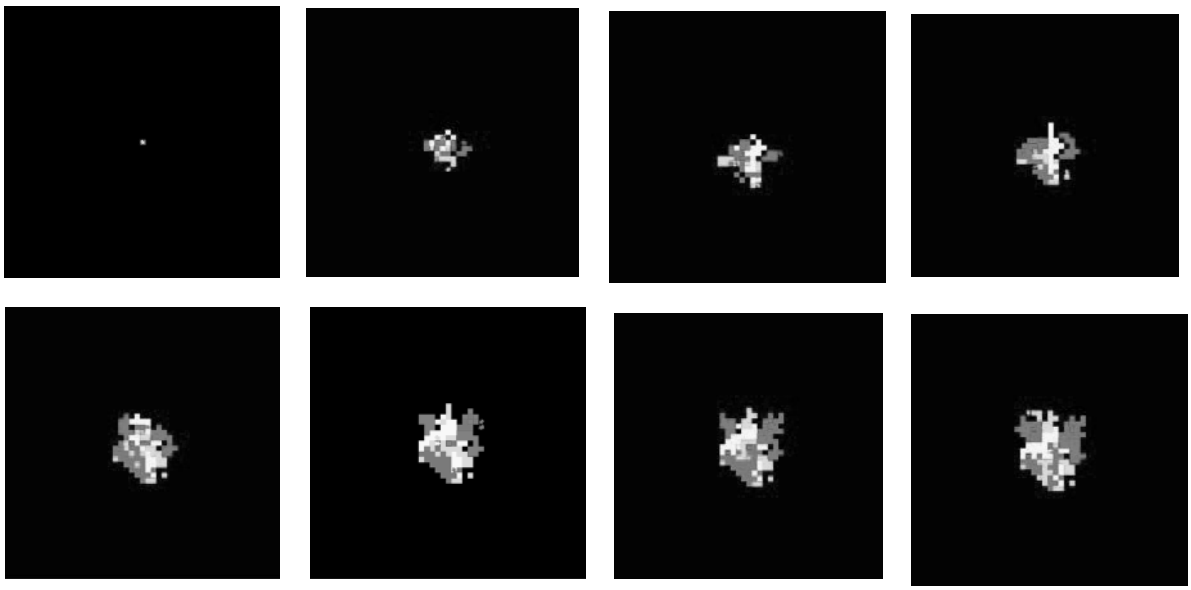


**Figure 3:** A sequence of steps for the first part of the model with the "gravitational wiggle".

An enhancement of this conceptual model was enhanced by the introduction of an environmental variation within the land by adding a third attribute to the land unit. This attribute was applied as a multiplier to the yield and was used to represent soil fertility. Orange patches would be defined by the user and used as an initial condition. These orange patches would then be assigned a value less than 1 and the wilderness patches would be assigned a value of zero as shown in Figure 4. The area is distorted by the soil fertility multiplier, but it can be seen that this distortion is neither regular nor predictable.

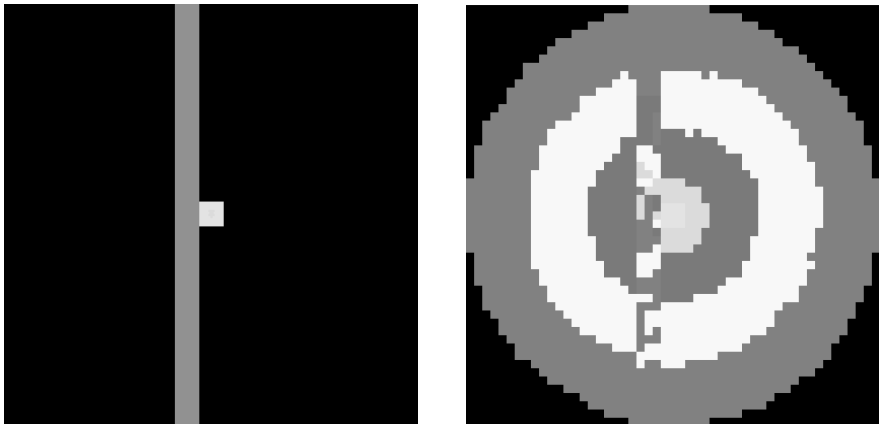


**Figure 4:** The user can define the area of land which is affected by soil infertility (orange) and run the model. The agents then mobilise and the area that is affected by the soil fertility is clearly shown in the final sequence.

As illustrated by Steadman's model (1998), it is possible to restrict the development of land within this model using natural barriers, such as a coast and a river. Two towns, Scarborough and Ipswich were chosen for their coastline and river respectively and each of these scenarios were created manually and saved as patches-own-variable and were broadly digitised from Google Maps (Google, 2010). As the model runs the agents were essentially turned away from the area of water and relocated to another patch successfully developing the land in the areas surrounding the water body as

shown in figure 5.

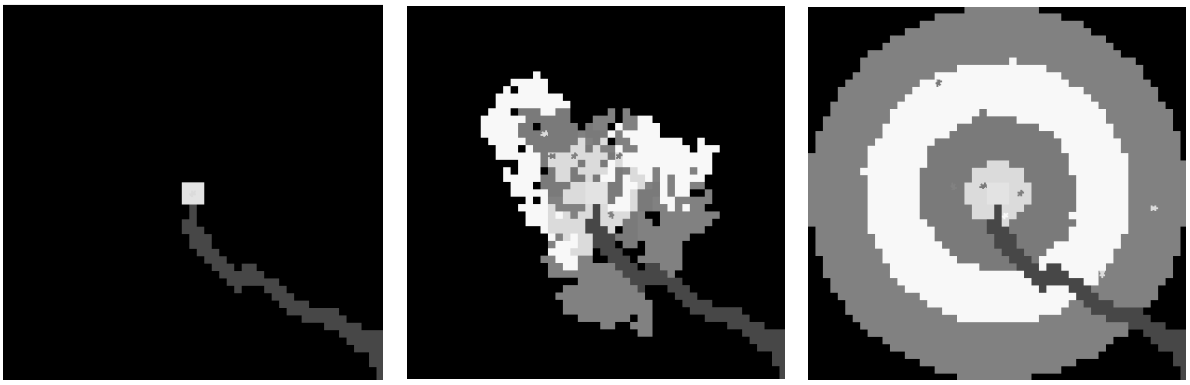


**Figure 5:** Sequencing for the first part of the model with the "gravitational wiggle".

**4. Conclusions and Future Developments**

From an educational perspective this research illustrates that agent based models do not need to be so reflective of real human behaviour and are investigations into complexity and its resulting patterns. It was encouraging that the core of the model retains the equation fundamental to von Thünen's original theory. The simple mechanism that illustrates the process of farms eventually operating at extreme margins (zero market profit) governed by land rent, allows users and further development to concentrate on altering the model's environment. This means changing von Thünen's original model from one of an isolated state on a homogeneous plane, to one where that state exists in a varied terrain and with possible variable conditions.

Current development of the model includes the introduction of barriers, shortest path algorithms, and transportation. These affect the variable d, distance to market. It is not difficult to envisage how the introduction of a railway station could eventually develop into satellite markets, and in turn into a network of towns. All the variables in the model need methodical tests for sensitivity and validation. Finally, further enhancements to the environmental variation multipliers could develop into epidemiological studies of crop pathogens or long terms effects of global warming.

**References**

Alonso, W. (1964). *Location and Land Use.* Cambridge, MA: Harvard University Press.

Argonne National Laboratory, (2009) *Recursive Porous Agent Simulation Toolkit (REPAST)*, (Software) available online at http://repast.sourceforge.net/

Box, P. and Sasaki, Y. (2003). Agent-Based Verification of von Thünen's Location Theory. *Journal of Artificial Societies and Social Simulation vol. 6, no. 2.* Available online at http://jasss.soc.surrey.ac.uk/6/2/9.html

Brown, D.G. 2006. Agent-based models. In H. Geist, Ed. *The Earth's Changing Land: An Encyclopedia of Land-Use and Land-Cover Change.* Westport CT: Greenwood Publishing Group, pp.7-13.

Google Maps, http://maps.google.co.uk/ [last checked 28.02.2010].

MIT Media Laboratory, (2006). *StarLogo 2.2* (Software). Media Laboratory and Teacher Education Program - MIT, available online at http://education.mit.edu/starlogo/

Otter, H. S., Van Der Veen, A. & De Vriend, H. J.( 2001). ABLOoM: Locational behaviour, spatial patterns, and agent-based modelling. *Journal of Artificial Societies and Social Simulation.*4 (4).

Parker, D. & Filatova, T. (2008). A conceptual design for bilateral agent-based land market with heterogeneous economic agents. *Computers, Environment and Urban Systems*, 454-463.

Resnick, M. (1998). *Turtles, Termites and Traffic Jams*. Fourth Edition. The MIT Press, Cambridge MA

Swarm Development Group, (1999). *Swarm* (Software). Available online at www.swarm.org. Swarm Development Group, Santa Fe, New Mexico.

Steadman, P. (1999). *Feeding the City: Von Thünen's Model of Agricultural Location*, Open University, Milton Keynes. Available online at http://www.casa.ucl.ac.uk/software/vonthunen.asp

The von Thunen Rings, http://www.prosper.org.au/wp-content/uploads/2008/08/thunen11-201x300.gif [last checked 28.02.2010].

von Thünen, J H, (1826). *Die isolierte Staat in Beziehung auf Landwirtshaft und Nationalökonomie*. Pergamon Press, New York. English translation by Wartenberg C M in 1966, P.G. Hall, editor.

Wartenberg, C M, 1966, *The Isolated State: an English Edition of Der isolierte Staat*. Pergamon Press, Oxford.

**Biography**

*Shaman Pottage is a postgraduate student at Birkbeck, University of London. Her research interest includes agent based simulation, environmental modelling, public health and geovisualisation.*

*Joana Barros is a lecturer in GI Science and the MSc in GI Science Programme Director at Birkbeck, University of London. Her research interests include computational models of geographical systems, agent-based simulation models of urban systems, as well as urban growth and change in developing countries.*

*Jez Nixon is a recent graduate of Birkbeck, University of London. He is interested in agent based simulation, web based GIS, and geovisualisation.*

*Alistair Cannell is a postgraduate student Birkbeck, University of London. He is currently employed by DigitalGlobe as a GIS specialist and he is interested in agent based simulation, web based GIS, and geovisualisation.*

# Modelling urban growth of Dhaka, Bangladesh

Sohel Ahmed[1], Glen Bramley[1]

[1]School of the Built Environment, Heriot-Watt University, Edinburgh, Riccarton, EH14 4AS
Tel. +44 (0) 131 451 4601/4605
Email: sja5@hw.ac.uk, G.Bramley@sbe.hw.ac.uk, www.sbe.hw.ac.uk

KEYWORDS: urban growth, Dhaka

## 1. Introduction

Dhaka, from a large town of fewer than 350,000 in 1951, has become a mega-city of over 11 million (Dewan and Yamaguchi 2009). By 2015, Dhaka's projected population of 19.5 million will fill most of the designated metropolitan area as a result of urban migration, extensions in the peripheries, and fresh urbanization (Islam 1999; Siddiqui 2004). Today it is evident that future urbanization in Dhaka presents enormous spatial and socio-economic challenges as new development must take place in fringe areas of mega-city Dhaka, despite being prone to flooding and containing very productive farmland. For an integrated urban planning strategy it is necessary to recognise, anticipate, measure, and understand urban dynamics and their consequences. Two appropriate spatial modelling frameworks conducive to developing countries have been used to comprehend dynamism of rapid urban growth and drivers influencing growth dynamics in the fast growing mega city Dhaka. Examples of near-future simulations incorporating spatial growth strategies from the model have also been presented.

## 2. Methodology and data preparation

There are certain reasons to choose these two modelling frameworks. The modelling frameworks are available in public domain, have friendly user interface and are transparent enough in their process of modelling when incorporating spatially and temporally explicit decision-making processes. Despite having no commercial value, these models have sufficient experiment potentials, in which a number of spatial conditions 'if . . . then' can be tested easily in a realistic way within a known degree of accuracy and with easier abilities to tailor to context.

For Dhaka, recently extracted spatial data on Land-Use and Land Cover Change i.e. LUCC (based on acquired RS imagery of 1960, 1975, 1988, 1999 and 2005 -validated by field survey and other secondary sources) has been used to understand historic growth trend and trajectories of urban Dhaka (Dewan and Yamaguchi 2009).From literature studies it has been found that Dhaka started to experience rapid urban growth since the late 80s and it never declined since. An attempt to explore the growth dynamics using some spatial growth metrics (not reported here ) validated such deduction.

This research has attempted initially to find statistically significant factors influencing urban growth using binary logistic regression in two periods (1988-99 and 1999-2005) as a function of the independent variables in the Table 1. As the analysis primarily focuses on 1988 onwards, the study used the time slices for calibration and validation respectively before progressing into the future simulation excercises..

These are used in CLUE-s modelling framework[1] to generate urban growth probability. Apart from

[1] The *CLUE* (Conversion of Land Use and its Effects) modelling framework ,developed at Wageningen University , published its first application in 1996 . CLUE-S is specifically developed for the spatially explicit simulation of land-use change based on an empirical statistical analysis of location suitability combined with the dynamic simulation of competition and interactions between the spatial and temporal dynamics of land use systems- using high-resolution data in which each pixel only contains one land-use type. For more details about the CLUE-s model, see Verburg, P. H., W. Soepboer, et al. (2002).

this statistical approach, the study also employed an hybrid approach under DINAMICA EGO modelling framework[2] -incorporating statistical model in the form of logistic regression or weight-of-evidence approach along with cellular automata functions.
All the analysis and data preparation are done at a 30m×30m resolution, which is the original resolution of the land cover maps.

Floods become regular feature in rapidly urbanizing Dhaka during the monsoon season and therefore, elevation data can be very significant influencing growth of Dhaka**.** A common variable in economic models of land-use change is the distance to the nearest location of employment or facility which is calculated as proxy for the costs associated with travel (Cheng and Masser 2003; Verburg, Schot et al. 2004; Verburg, Kok et al. 2006; Braimoh and Onishi 2007). Advanced raster-based accessibility measures (cost-distance model ) are calculated in travel times - using ArcGIS Desktop Model Builder™. Samples of the resultant variables can be seen in figure (1). Enrichment factor[3] has been used as neighbourhood variables for quantifying and analysing neighbourhood characteristics and was most influential at the 10th neighborhood.

**3 Results and discussion**

From the table 1, it can generally be deducted based on the wald statistics[4] and odds ratio that the major determinants of urban development have changed in time: from distance to the old CBD to closeness to other minor city centres; proximity to major roads is no more the dominant one and physical condition of the sites like elevation has become less significant. Probability (log –odds) of being urban decreases by 0.011 with every minute away from old CBD in the former period – characterizing the decentralized, polycentric suburbanizing trend in the metropolitan Dhaka area which becomes less important a factor for the later period (evident through Wald statistics). But not all results lead to good explanation or not as expected / influential. For example, planning variables, particularly urban fringe areas for accelerated development, has not resulted as significant for the time 1988-99. This is due to the fact during that time, there were still room for consolidation within the inner city areas and such has been reflected by the significant influence of factors like travel time to inner city facilities namely CBD, commercial and industrial areas. Accessibility to major roads is an important determinant of urban land change, the variable did contribute significantly to the model as expected between 1988-99 i.e. areas near to major road are 0.171 times more likely to turn urban than those minutes away from these roads. But this variable didn't predict much growth in 1999-2005 as data on road at the expansion area are not available. The prospect of eastern bypass has proliferated much growth at the eastern side of the city during this time with closer distance to the new vibrant minor city centres nearby, even though these areas are not free from flood risk.

As can be seen from figure 2 and 3, probability maps are generated at two levels: only using the driving factors and then taking into account partial neighbourhood effects. The later appears to be better able to predict the spatial pattern of urban development. In both time periods, areas at mid-north, west, central areas and areas at south-east showed higher probabilities to change to urban. Probability map for 1999-2005 started to provide weights to eastern areas which are experiencing

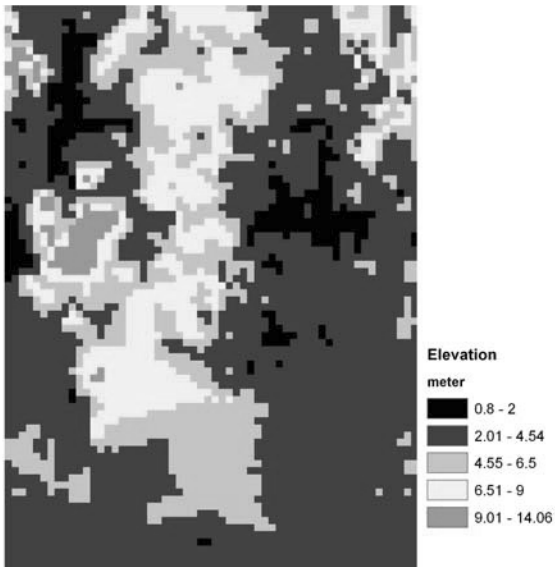
---

[2] It is developed at the Center for Remote Sensing of the Federal University of Minas Gerais (CSR-UFMG), Brazil. It has been designed as a general purpose dynamic modeling software, -implementing the most common spatial analysis operators available in commercial GIS, plus and a series of complex spatial algorithms for the analysis and simulation of space-time phenomena . For more details about the DINAMICA EGO model, see Almeidaa et al., 2003; Soares-Filho et al., 2002

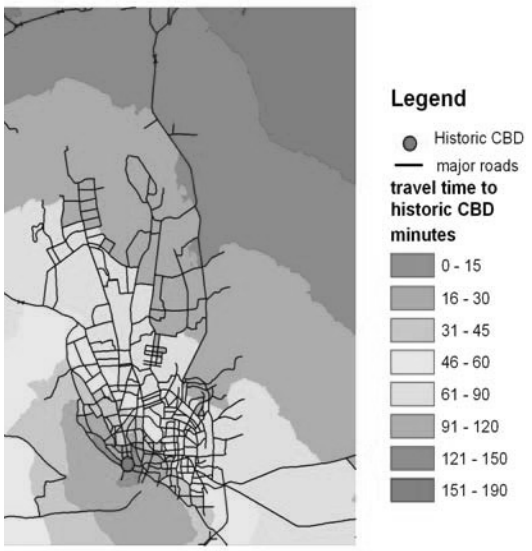
[3] Enrichment factor, as coined by Verburg et al., 2004, is a measure that characterises the over- or under-representation of different land-use types in the neighbourhood (of a specific grid cell) by comparing the amount of occurrences of a particular land-use type in the vicinity of a specific location as relative to the volume of occurrences of that land-use type in the study area in total.

[4] The Wald statistic is the squared ratio of the unstandardized logistic coefficient to its standard error.

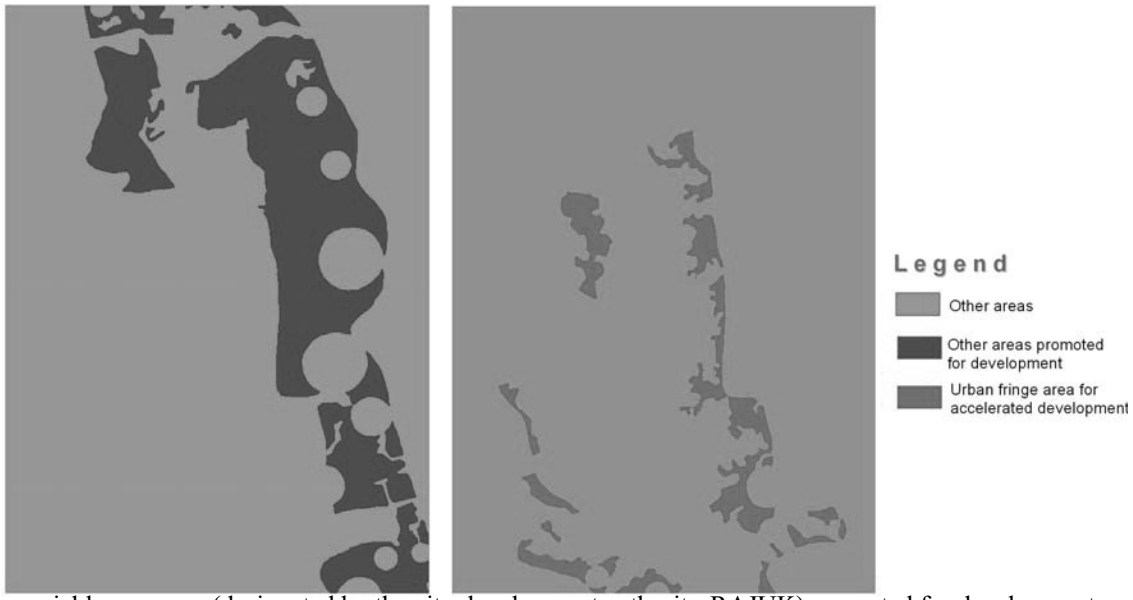
rapid conversion in recent years.




Elevation layer




Travel time in minutes to historic CBD




Planning variables as areas (designated by the city development authority-RAJUK) promoted for development and fringe areas for accelerated development by the city

**Figure 1:** Samples of spatial layers used for calculating urban growth probability

**Table 1 : Logistic regression analysis of driving factors with urban land 1988, 1999 and 2005 as dependent variable**

| | | | | | | |
|---|---|---|---|---|---|---|
| | $\beta$ | $\exp(\beta)$ | Wald statistic | $\beta$ | $\exp(\beta)$ | Wald statistic |
| Travel time to commercial centres (in minutes) | -.029 | .971 | 1787.262 | -.015 | 985 | 444.619 |
| Travel time to historical CBD | -.011 | .989 | 2256.089 | -.006 | .994 | 821.670 |
| Travel time to industrial hubs (in minutes) | -.018 | .982 | 1117.147 | -.018 | .982 | 1012.448 |
| Travel time to minor city centres (in minutes) | -.024 | .976 | 904.187 | -.036 | .965 | 2586.417 |
| Travel time to major roads (in minutes) | -.171 | .843 | 8274.910 | -.036 | .965 | 1411.041 |
| Elevation in metres | .239 | 1.279 | 3887.087 | .131 | 1.139 | 1650.640 |
| Planning variable (urban fringe acceleration) | * | * | * | -.261 | .770 | 104.517 |
| Planning variable (promoted areas for development) | -.104 | 1.068 | 4.309 | -.104 | .901 | 55.056 |
| Constant | 1.369 | 3.933 | 2201.911 | .965 | 5.771 | 1300.033 |
| Adjusted $R^2$ ( Nagelkerke) | | | .353 | | | 0.152 |
| ROC | | | 0.865 | | | 0.703(NP) and 0.702 (P) |
| Reported urban % matched at classification table | | | 28.6 (84.2% overall) | | | 19.1 (72.6% overall) |

* not significant at 0.05 level; Rest of the results reported are significant at 0.05 level $\beta$ **=Unstandardized coefficients ;** $\exp(\beta)$ **= odds ratio**

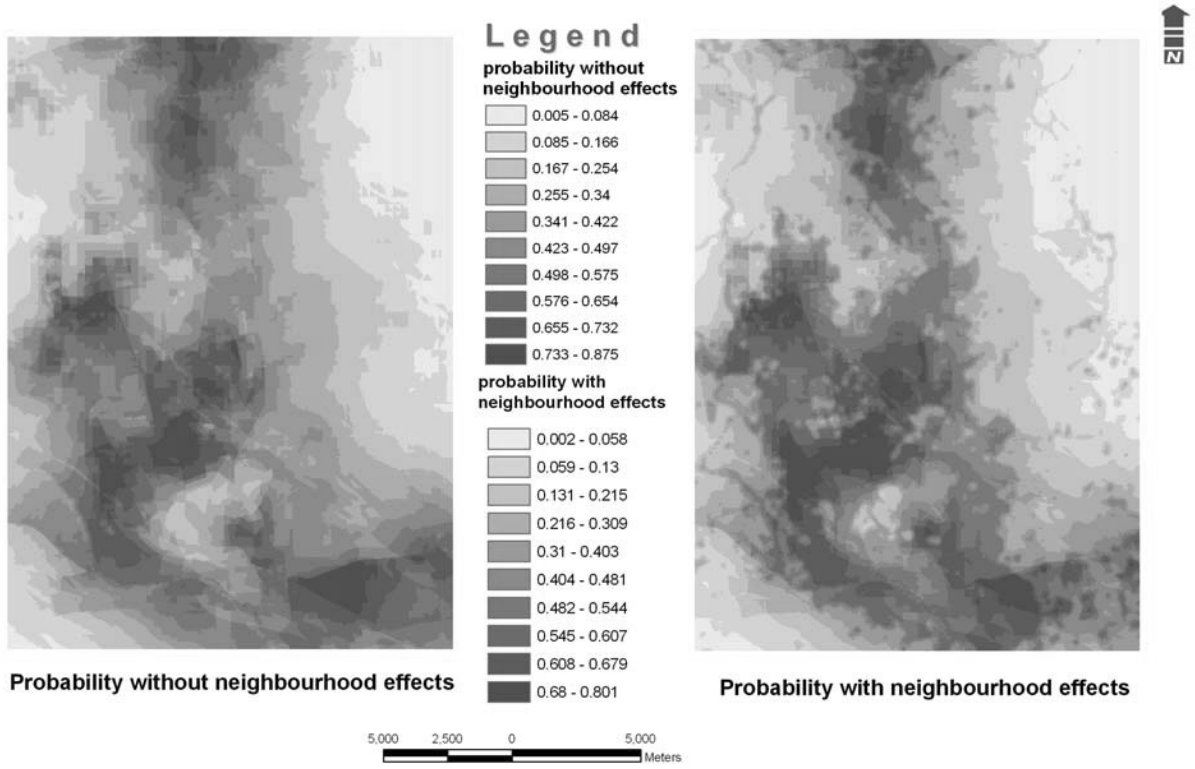



**Figure 2: probability maps for 1988-99 with and without neighbourhood effects generated in CLUE-s and ArcGIS**

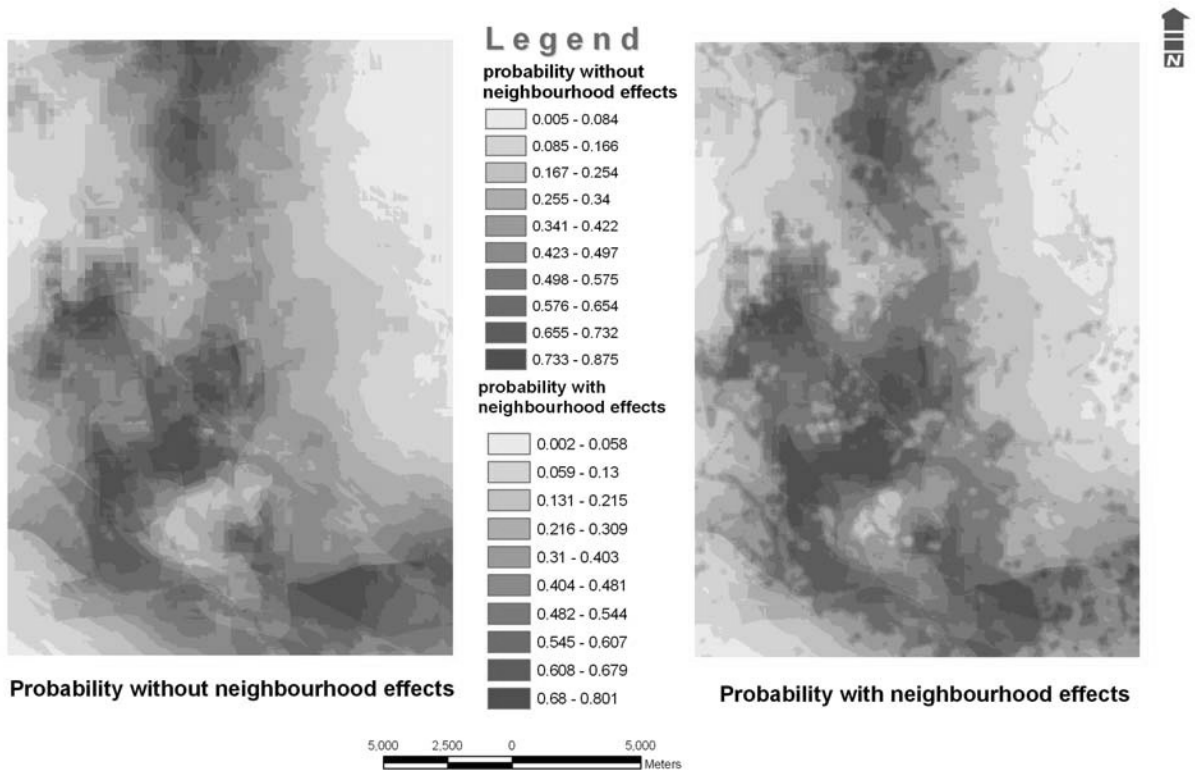



**Figure 3:** Probability maps for 1999-2005 with and without neighbourhood effects generated in

CLUE and ArcGIS

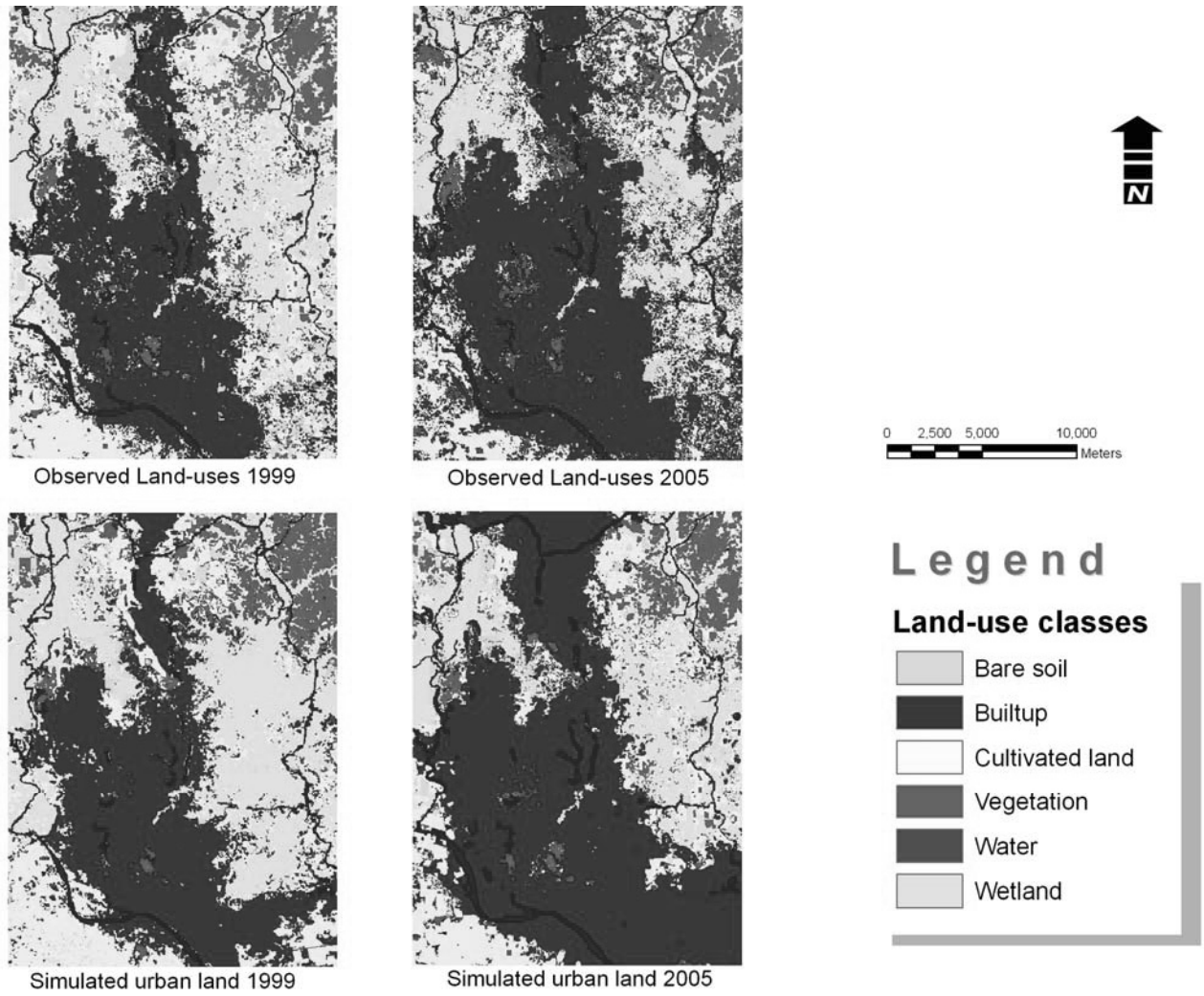



**Figure 4 :** Observed and simulated land-use for 1999 and 2005 in CLUE-s

At this stage urban land has been simulated in CLUE-s using observed land-use patterns of 1988 and 1999 with enrichment factors and relevant regression parameters, The best results based on calibration and validation exercises (not included here) were presented with the observed land-use change of 1999 and 2005 in figure 4. There are extensive hinterlands at the northern extensions of the city and that in the eastern regions that are not at all reflected within the simulation results. The results when incorporating neighbourhood effects looks to cover the infill mechanism well .But these seem to fail to include the chance events of other land-uses becoming urban. To capture spontaneous processes, a hybrid approach -incorporating either weight of evidence approach or logistic regression as well as cellular automata techniques have been employed in DINAMICA EGO using same datasets. For the weight-of-evidence method, transition probabilities governing changes in land use (as functions of a variety of socio-economic and infrastructural factors) are measured through spatial correspondences akin to the methods of map overlay (Almeidaa, Batty et al. 2003).

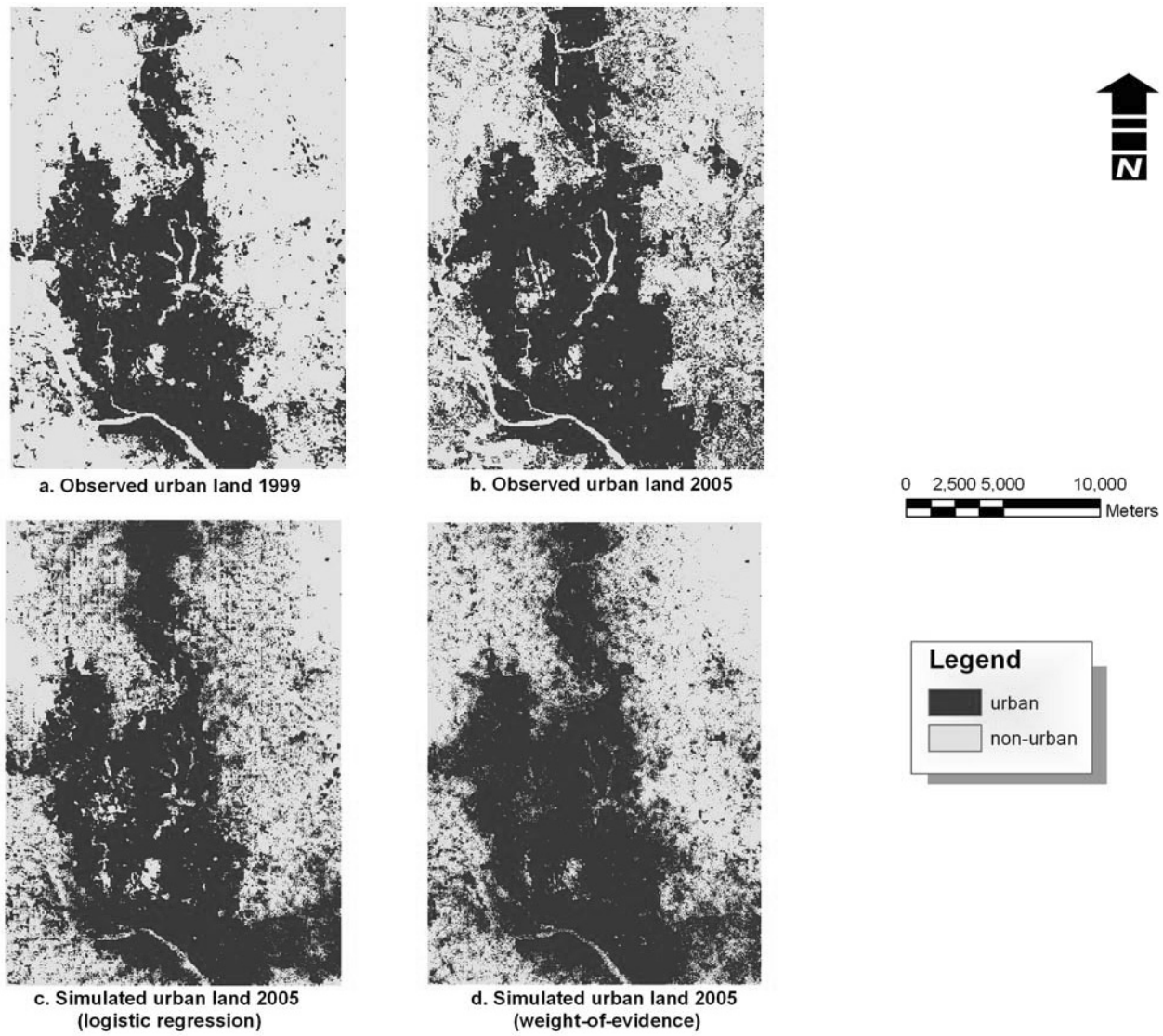



**Figure 5:** Observed urban land for 1999(a), 2005(b) and simulated urban land 2005 using logistic regression (c) and weight-of-evidence (d)

While checking them visually and through validation statistics comparing with a neutral model the results from weight-of-evidence probabilities performed better (figure 5).

This is a good indication that the model has sufficient predictive power to make near future simulation, since random model has similarities with recent urban changes in Dhaka in terms of randomness and the model came close to those observations. Therefore a prediction model using weight-of-evidence has been chosen to make near future scenarios in the next step.

In figure 6, the business-as-usual scenario has been found realistic to some extent with the present trend as Dhaka has still been consolidating in the inner core areas, where most employment opportunities exist and population density have been highest (RAJUK 1997). Although having all inner open spaces to be taken over is highly unlikely, this does provides an eye opener what might happen if present practice of weak planning and development control continues.

For this research, principal spatial development strategy as outlined in DMDP structure plan (1995-2015), has been adopted to examine near-future scenarios for Dhaka. For the 2$^{nd}$ scenario, city's existing natural drainage system and open spaces (as of 2005) has been kept constant i.e. these lands can't be converted to urban land to scenarios for next 10 (2015) and 20 years (2025). For the 3$^{rd}$ and

last one, designated places as flood retention ponds, flood flow zones and areas with high agricultural values are added to the constraints put in place already.

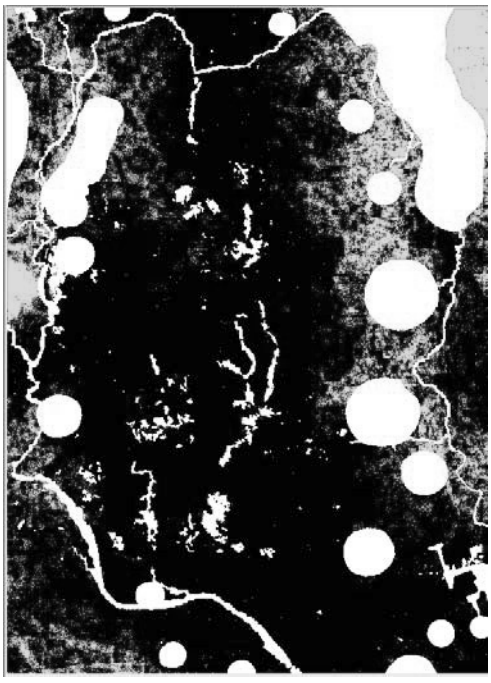
Bringing planning constraints into the simulations, changes the urban growth pattern which can be noticed by bare eyes. These scenarios assist in getting an impression where urban growth of Dhaka should march in 10 or 20 years if rigid planning and development control practices are put on places from now on.

**Dhaka 2015** **Dhaka 2025**

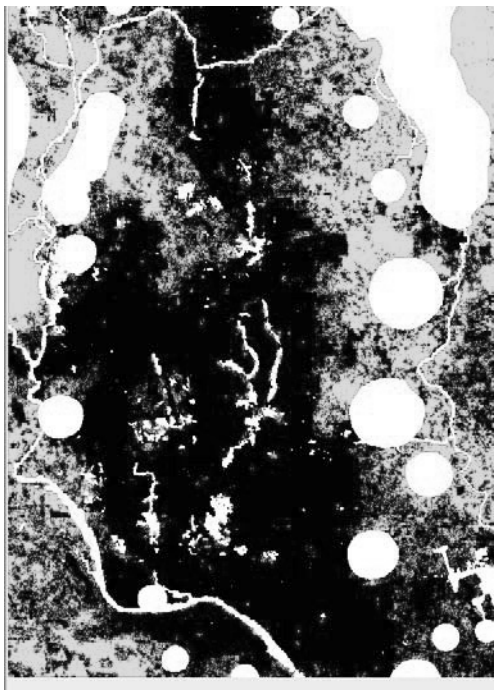
**a.**

**b.**

**c.**



**d.**


**e.**


**f.**

**Figure 6:** Business-as-usual scenario for 2015(a) and 2025(b); planning restriction scenario by keeping rivers and other existing inner open spaces intact for 2015 (along with restricting growth in flood flow zones and high-value agricultural lands for 2015 (c) and 2025 (d); planning restriction scenario with reduced growth rate identical to 1988-99 period for 2015 (e) and 2025 (f);

**4. Acknowledgements**
This research is funded by ORS-James Watt scholarship at Heriot-Watt university. Besides the researchers are indebted to Ashraf Dewan for sharing LUCC data for Dhaka.

**References**

*Almeidaa, C. M., M. Batty, et al. (2003). "Stochastic cellular automata modeling of urban land use dynamics: empirical development and estimation." Computers, Environment and Urban Systems **27**: 481–509.*

*Braimoh, A. K. and T. Onishi (2007). "Spatial determinants of urban land use change in Lagos, Nigeria." Land Use Policy **24**: 502–515.*

*Cheng, J. and I. Masser (2003). "Modelling urban growth patterns: a multiscale perspective." Environment and Planning A **35**: 679 -704.*

*Dewan, A. M. and Y. Yamaguchi (2009). "Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization." Applied Geography.*

*Islam, N. (1999). Dhaka city: Some general concerns. Asian Cities in the 21st Century: Contemporary Approaches to Municipal Management. Manila, Philippines, Asian Development Bank (ADB). **Vol. 3: Reforming Dhaka City Management:** 71-82.*

*NPS-ROMN (2008). Travel Time Cost Surface Model. Fort Collins, Colorado, Rocky Mountain Network-National Park Service.*

*RAJUK (1997). Dhaka Metropolitan Development Plan for Dhaka city (1995–2015) Vols. I & II. Dhaka, Rajdhani Unnayan Katripakkha (RAJUK).*

*Siddiqui, K. (2004). Megacity Governance in South Asia: A comparative study. Dhaka, Bangladesh, University Press Limited (UPL).*

*Verburg, P. H., K. Kok, et al. (2006). Modeling Land-Use and Land-Cover Change. Land-Use and Land-Cover Change: Local Processes and Global Impacts. E. F. Lambin and h. Geist. Verlag Berlin Heidelberg, Springer.*

*Verburg, P. H., T. C. M. d. Nijs, et al. (2004). A method to analyse neighbourhood characteristics of land use patterns. Computers, Environment & Urban Systems **28**(6): 667 -690.*

*Verburg, P. H., P. Schot, et al. (2004). "Land use change modelling: current practice and research priorities." GeoJournal **61**(4): 309 - 324.*

*Verburg, P. H., W. Soepboer, et al. (2002). "Modeling the Spatial Dynamics of Regional Land Use: the CLUE-S Model." Environmental Management **30**(3): 391-405.*

## Biography

**Sohel Ahmed** is PhD student in Urban Studies at Heriot-Watt University, Edinburgh.

**Glen Bramley** is Professor of Urban Studies at Heriot-Watt University in Edinburgh, where he leads a substantial research programme in housing and urban studies. Recent work is focused particularly on planning for new housing, housing need, home ownership, neighbourhood change, urban form, quality of life, poverty and the funding and outcomes of local services. His publications include *Key Issues in Housing* (Palgrave 2005), and *Planning, the Market and Private House building* (UCL Press 1995).

# Modelling of Crime Hotspot on Complex Street Network

Chen Peng, Yuan Hongyong, Ni Shunjiang

Centre of Public Safety Research, Department of Engineering Physics, Tsinghua University, Beijing, 100084
Tel. 8610-62792894 | Email: p-chen07@mails.tsinghua.edu.cn

KEYWORDS: network; agent; hotspot

## 1. Introduction

Since the phenomenon of crime concentrating in identifiable places was noted by Brantingham and Brantingham (Brantingham & Brantingham, 1982) and then was investigated by Sherman, et al (Sherman, Gartin & Buerger, 1989), it had been studied by many people. One of the opinions considered that the crime was intimately associated with the physical environment in which it occurred, therefore, crime will be clustered intensely in certain places.

In reality, many property crimes like robbery occurred on outdoors. Because people perform their routine activities between homes, workplace, recreation sites and shopping centre via the road or street, the topology of the road or street will influence the intersections of people. In past years, some researches had shown that the configuration of street networks would influence crime's placement (Bevis & Nutter, 1977; Beavon, 1984; Brantingham & Brantingham, 1993), so, in this paper, a study about simulating crime (robbery) hotspot in street network with agent based model was presented. In following sections, the framework of model was introduced and then the findings of the simulation were presented.

## 2. Framework of the model

### 2.1 Agent

In the model, three kinds of agents were setup to represent the role of civilians, police officers and motivated offenders according to the principle of routine activity theory (Cohen & Felson, 1979). For the agents, civilians make living by working, but for offenders, they maintain their lives by committing crimes (Nuňo, Herrero, & Primicerio, 2008). The agents moving behaviours were simulated according to statistic distribution given by Gonzalez (Gonzalez, Hidalgo, & Barabasi, 2008). Specifically they include three steps: travelling, waiting and returning.

(*i*) Travelling: for each individual agent $h$, initially it was located at node $i$, at each time step, it will leave the node $i$ with probability $p_{jump}$ and enter into a neighbour node $j$, then go to step (*ii*); therwise, individual $h$ will remain in node $i$.

(*ii*) Waiting; when individual $h$ enters into a neighbour $j$, it will be assigned a waiting time $t_w$ that is drawn from a heavy-tailed distribution $p(t_w) \sim t_w^{-(1+\lambda)}$, where $\lambda>0$ and $1 \le t_w \le T_{max}$; then the individual $h$ will wait for $t_w$ time steps in node $j$ before it travels again.

(*iii*) Returning: After waiting for $t_w$ time steps in node $j$, the individual $h$ will return back to node $i$ with probability $p_{back}$, and then start from step (*i*) again at next time step; otherwise, it will travel to another node $k$ that is selected according to the rules in step (*i*).

For offender agents, their moving behaviours are a little different from police and civilian agents.

When they have some properties, they act as civilians, but once they run out of their funds, they will random walk to search for suitable targets and then commit crimes to obtain property for maintaining their lives.

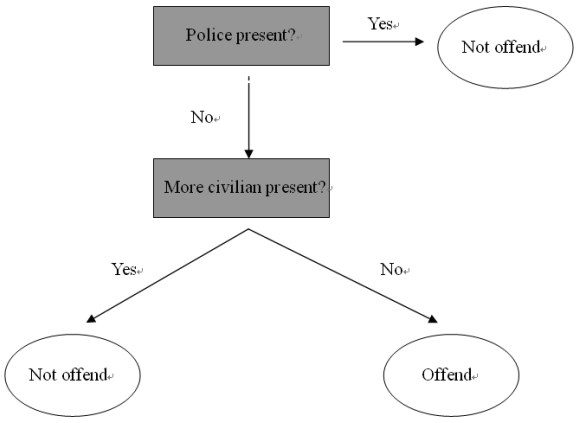



**Figure.1** Decision process of offender agents

Offender agents have a decision process to evaluate the circumstance and decide when to commit crime (see Figure 1). When offenders intersect with police, no crime will occur however the targets were suitable. When there was no police present on node, offenders will evaluate how many civilian agents present on node because more civilians will act as informal guardians (Groff, 2007). Thus, in this condition, whether a crime would occur was determined by the formulation (1):

$$G = (N_C - N_O) \tag{1}$$

If $G < 0$, then there was a lack of capable civilian agents so a crime would occur

If $G \geq 0$, then there were capable of civilian agents so a crime would not occur

Where $\boldsymbol{G}$=guardianship, $\boldsymbol{N_C}$=number of civilian agents at node, $\boldsymbol{N_O}$=number of potential offender agents.

### 2.2 Landscape

The landscape where the agents perform their routine activities was made up of nodes and paths. The nodes represented the locations or facilities in which agents live or work. The paths were the connections that connecting the nodes. Some previous researchers made study about the structures of road or street network and they found the nodes connections were distributed in power law and had small world characteristic (Crucitti, Latora, & Porta, 2006).

In this work, two types complex network - WS and BA scale free - were used to represent the street work in real life. WS network is a kind of mathematic graph, and it could be produced according to an algorithm (Watts, Strogatz, 1998). Firstly, starting from a ring lattice with $n$ vertices and $k$ edges per vertex, then each edge are rewired at random with probability $p$, $p$ lies between the interval of 0 and 1, When $p=0$, the graph is regularity; while $p=1$, the graph is disorder.

BA scale-free network is another small world network whose physical features are different from WS. In BA scale-free network, the nodes' connections follow a power law distribution, which means some common vertices have higher connections, but most of the others have few connections. The highest-degree nodes are often called "hubs". Barabási and Albert (Barabási & Albert, 1999) introduced a method named preference attachment to construct BA scale free network: new nodes are added to an initial pre-existing network with $m$ nodes, connected to each of the original nodes with a probability proportional to the number of connections each of the original nodes already had.

Generally, the characteristics of complex network were evaluated by two parameters: clustering coefficient and average shortest path. The clustering coefficient was defined as formulation (2). The higher parameter means the nodes was intensely connected with its neighbours. Comparatively, average shortest path was used to measure the connectivity of the network. Smaller parameter means the node in network could be reached from any other ones via limited connections.

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{2}$$

Where $k_i$ is the neighbours of the $i$th node, $E_i$ is the number of connections of the $i$th node.

### 2.3 Parameters

The parameters and corresponding rationales were presented in Table 1.

**Table 1.** Modelling parameters

| *Parameter* | *Rationale* |
|---|---|
| Number of nodes=10000 | Large enough to construct a complex network |
| Initial connections in ring lattice ($k$)=4 | In most cases, intersections are crossed by two roads thus produced four connections. |
| Rewiring probability ($p$)=0.01 | The features of small world network is most significant in this probability level |
| Initial nodes in network ($m$)=2 | The minimum nodes that are desired to constructed BA scale-free network. |
| Number of civilians=10000 | Every node has an owner |
| Number of offenders=100 | the number of offenders in population is quite small |
| Number of police=100 | Maintain the same level with offenders |
| $P_{jump\_day}$=0.9 | The probability of agent leaving node at day time |
| $P_{jump\_night}$=0.01 | The probability of agent leaving node at night time |
| $T_{max}$=72 | Every time step represent 10 minutes, so a half day equals to 1440 minutes |
| $\lambda$=1.6 | Reference from literature |
| Time step=52560 | Model one year |

## 3. Findings

The model was simulated for 100 times and the results were averaged from the simulation. The statistic parameters of the networks and simulation findings were shown in the Table2. According to the table, the total amount of crime in both networks were almost same, however, they had different spatial distribution patterns. In BA network, the number of nodes where crime had occurred was smaller than those in WS network. So, it was demonstrated that crime inclined to be more clustered in BA network than in WS network.

**Table 2.** Statistics from the simulation

| | *WS small-world* | *BA scale-free* |
|---|---|---|
| Clustering coefficient | 0.468 | 0.00218 |
| Average shortest path | 45 | 6 |
| Total intersections | 11710 | 15418 |
| Total robberies | 7294 | 7296 |
| Nodes with robberies | 1202 | 1089 |

The nodes in both network was listed as a series according to the number of crime had occurred, then the cumulative crime level of these nodes was calculated. The result was displayed in Figure 2. The findings indicated that crime incidents in BA network concentrated more intensely than in WS network. For example, in top 200 nodes, the cumulative crime level in BA network approached 55%, while this proportion in WS network was only 38%.

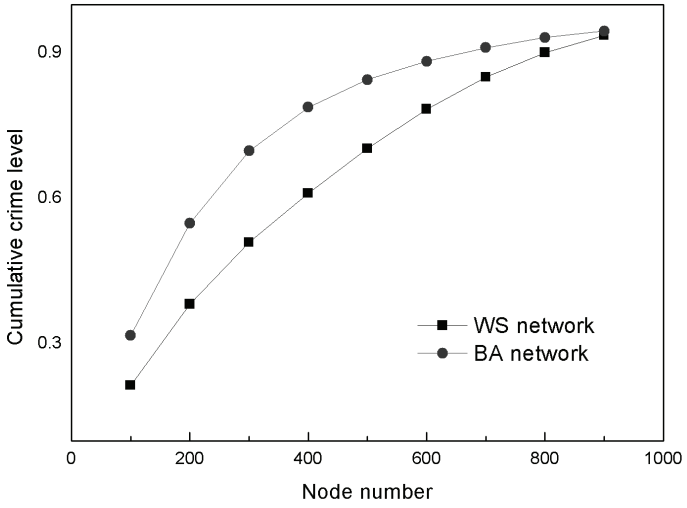



**Figure 2** Cumulative crime level in nodes in both networks

The difference of crime spatial patterns in both networks was due to their topology structure. According to Table 2, the average shortest path in BA network was about 6, which meant agents could reach other nodes via limited nodes, therefore, in this network agents could have more intersections (see row 3 in Table2). But for WS network, this parameter was 45, so the agents in this structure had fewer intersections than in BA network. Besides, the clustering coefficient of BA network was much smaller, which means most nodes in this network had fewer connections but only a small part of nodes was connected more intensely. So, on these nodes, the agents had more opportunities to intersect with others so that the crime level on these nodes was very high (Figure.3).

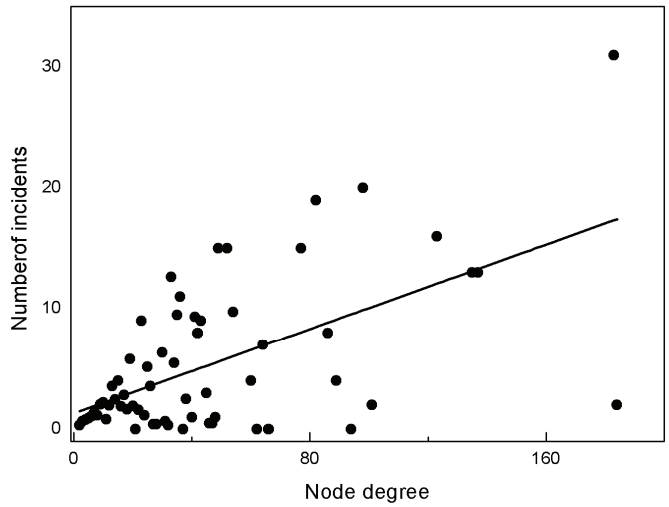



**Figure 3** Relationship between robbery number and node connections in BA network

## 4. Conclusion

This study investigated crime spatial patterns in two street topology structures. The result demonstrated that the network topology has significant influence on crime placement. Crime appears to be more clustered in BA network for little average shortest path and clustering coefficient. This finding informs that some places where more people visit will create more opportunities to intersect then produce higher crime risk. So strengthening monitoring or police patrolling in these places is a suitable way to prevent crime for decision agencies.

## References

Barabási A, Albert R (1999) Emergence of scaling in random networks. *Nature* 286, 509

Beavon D J K (1984) Crime and the Environmental Opportunity structure: The Influence of Street Networks on the Patterning of Property Offenses. Master's thesis, Simon Fraser University. urnaby. British Columbia, CAN.

Bevis C, Nutter J B (1977) Changing Street Layouts to Reduce esidential Burglary. Paper presented at the annual meeting of the American Society of Criminology, Atlanta, GA.

Brantingham P L, Brantingham P J (1982). Mobility, notoriety, and crime: A study of crime patterns in urban nodal points. *Journal of Environmental Systems* 11:89–99

Brantingham P L, Brantingham P J (1993) Nodes, paths and edges: considerations on the complexity of crime and the physical environment. *Journal of Environmental Psychology* 13: 3-28.

Cohen L E, Felson M 1979 Social change and crime rate trends: a routine activity approach. *American Sociological Review* 44:588–608

Crucitti P, Latora V, Porta S (2006) Centrality measures in spatial networks of urban streets. *Physical Review E* 73: 1-5

Gonzalez M C, Hidalgo C A, Barabasi A L. Understanding individual human mobility patterns. *Nature*, 2008, 453 (7196): 779-782

Groff E R (2007) Simulation for theory testing and experimentation: An example using routine activity theory and street robbery. *Journal of Quantitative Criminology* 23:75-103.

Nuňo, J.C, Herrero, M.A, Primicerio, M (2008). A triangle model of criminality. *Physica A*. 387: 2926-2936

Sherman L W, Gartin P R, Buerger M E (1989) Hot spots of Predatory crime: routine activities and the criminology of place. *Criminology* 27(1): 27-56

Watts D J, Strogatz S H (1998) *Nature* 393**,** 440

## Biography

*Chen Peng is a Ph.D student in Center for Public Safety Research, Tsinghua University. His interest is simulating spatial-temporal crime hot spot with agent based modelling. He received his MS and BS degrees from Dalian University of Technology.*

# Development and Application of a Probabilistic Time-Activity Model

Linda Beale[1], Duncan Whyatt[2], Federico Fabbri[1], Gemma Davies[2] David Briggs[1]

[1]Imperial College London, St Marys, Praed Street, London, Paddington, W2 1PG
Tel. 020 7594 3348 | Email, l.beale@imperial.ac.uk

[2]Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ

KEYWORDS: exposure, epidemiology, networks, GPS, journey to school

## 1. Introduction

In spatial epidemiology 'place of residence' remains widely used as the individual's location throughout the period of analysis. Using this location in space and time individuals exposures to various environmental factors are modelled. A number of studies have demonstrated that personal monitoring reveals differences in environmental levels, however, such data collection is time consuming and expensive. Individual exposure can vary depending on circumstances such as the amount of time spent indoors and outdoors, respectively (Jarup 2004). Furthermore, environmental pollutants and exposures vary spatially and temporally with for example, different levels during day and night (Briggs 2005). Increasingly geo-coded health data, population distribution and environmental levels are being gathered. Spatial epidemiology approaches to evaluate such data are improving in part with the increased use of Geographical Information Systems (GIS). All of which suggests an ever increasing need to more effectively model population exposure to help better understand people's different exposures in both space and time.

This paper describes the development and application of a probabilistic time-activity model for exposure assessment. Using this GIS-based model we can simulate an individual's daily journeys and evaluate daily exposures. Model output is compared with actual journey data captured by school children using GPS and mobile phone technology in an independent study. Initial results show that the model is able to effectively model children's travel mode for local journeys, however, more calibration is required to differentiate between the use of car and bus on longer journeys to school.

## 2. Background

Although in general people spend a relatively small proportion of their time travelling, travel environments are often relatively polluted. In the UK, research has shown that around 70% of particulates in urban areas are accounted for by traffic emission and road dust (Harrison et al 1997). Journey-time exposures may, therefore, comprise a disproportionately large amount of total personal exposures to ambient and in particular, traffic-related air pollution.

There are some sources of time-activity data but these tend to be aggregated data with little spatial and temporal detail such as the UK census origin-destination (OD) data. In other cases they may take the form of localised traffic or pedestrian counts. Increasingly research is taking advantage of GPS technology and in particular using mobile phones to track individual's movements (Walker et al 2009). Whilst these are effective in the spatial dimension, they are of little use to study longer term effects such as potential health outcomes to environmental factors in different locations since they require that data is collected each time analysis is required. The ability to simulate subgroup population movement would be a valuable resource for spatial epidemiological research.

Spatial epidemiology can be an effective method of investigating the links between the environment

and health. Advances in both GIS and statistics have significantly contributed to this area of research, however, in most cases the location of any individual is taken to be their home residence (e.g. contaminated groundwater plumes: Ball et al. 2008, landfill sites: Elliott et al., 2001; cokeworks: Aylin et al., 2001, roads: English et al., 1999). Many people spend much of their day in a different location e.g. at place of work, and are subjected to many exposures in their daily travel. Using spatial analysis and increasingly available data on population time activity, research should aim to quantify the importance of actual location as opposed to home location in spatial epidemiology.

## 3. Method

Time-activity surveys typically contain information on the type of location (e.g. home or work), in some cases occupancy times, but very rarely include detailed spatial data on journey routes. Combining probabilistic time-activity behaviour with route modelling in GIS provides the opportunity to model individuals' paths through the environment.

The methodology described here enables time-activity patterns to be modelled on the basis of variable levels of data. The model is probabilistic, and designed to extract the maximum amount of information possible from any available data. The model simulates time-activity patterns as a series of trips, with intervening 'static' periods at selected locations. Journey and occupancy times are modelled on the basis of population distributions from existing time-activity surveys or national statistics. The resulting simulations can be used to assess potential exposures of specific subgroups of people to environmental hazards (e.g. air pollution, noise, traffic accidents) and to compare risks or other indicators under different management or policy scenarios.

Time activities are modelled at the individual level with a set of time activity characteristics for the population or for a population subgroup, with its own, unique set of time activity characteristics. These individual routes can then be combined to make population distributions of exposure or exposure distributions for population subgroups e.g. schoolchildren. Such a probabilistic time-activity model could be used to enhance any existing time-activity data or simulate complete time-activity patterns at the individual level, as a basis for exposure modelling.

The model has been programmed in ESRI ArcGIS using network analysis tools. To fully calculate an individual's activities in the model, data is required on the start and end location of each trip (X/Y), the start and end times of each trip, trip mode, travel speed, route and activity type (i.e. activities undertaken at the destination). Using these data the model is able to reproduce the time activity sequence. If any of these parameters are missing from the time-activity database, however, they can be imputed probabilistically on the basis of available statistical information. These statistical data comprise generalised distributions for the study population. Time-activity modelling can thus be performed using a mixture of individual level and statistical (aggregated) data from time-activity surveys. So for example, time activities can be defined in terms of the percentage of the population at each of the selected destination types, for each hour of the day or using distance-decay functions for each of the required travel modes e.g. by car, bus, walking etc.

Routes are selected on the basis of the least-cost route by default within the GIS; this is the shortest route in terms of distance. Route cost can, however, be defined in terms of another form of impedance e.g. on the basis of the shortest travel time or easiest route. This method has been demonstrated to model wheelchair accessible routes (Beale et al. 2006) where an impedance value was calculated for all network segments using additional data on a number of factors that either hindered or improved wheelchair accessibility. In the probabilistic time activity model impedance is calculated using road type, speed and arc length

## 4. Outcomes

Using data collected as part of a study into the impact of traffic-related air pollution on the journey to School (Walker et al, 2009), daily time activities diaries were generated for school children in the

Lancaster area. This dataset provided 31 unique start points, one for each child, and one common end point, the school.

To date, this work has focussed on validating the models selection of travel mode to school. In the near future we also plan to validate model route selection, through comparison with actual routes taken by school children (captured using GPS technology). Ultimately, these data will be used to compare the modelled exposures that we would obtain using 1) home location only 2) typical known route and 3) modelled route. This research will attempt to quantify exposure misclassification that may be introduced into small scale spatial epidemiological analysis when the population is assumed to be static.

The schoolchildren's travel mode was modelled using a probability model based on national travel survey data (Cronberg et al. 2007). A number of potential routes between the known origins and destination were modelled and using the distances of these routes (by mode) a journey mode was estimated. Figure 1 shows a comparison between predicted and actual travel mode by distance (n=31, $r^2 = 0.65$). The results show that at distances of <2.5km the model is reasonably efficient (60% of travel modes correctly predicted). At greater distances, however, the model does not effectively differentiate between car and bus (46 % correctly predicted). The probability model used national travel to school data from the National Travel Survey 2006, however, improved data from the National Travel Survey, 2002-2006 will now be incorporated into the model. This combines data collected over a number of years and allows for improved breakdown of subpopulations (e.g. between ages of schoolchildren).

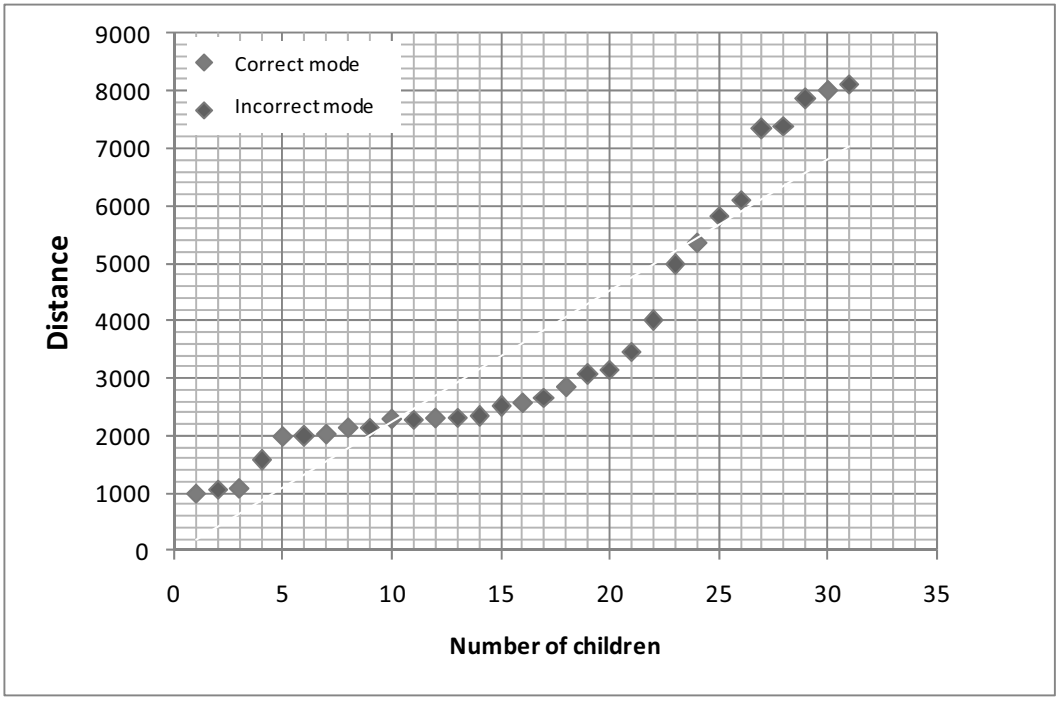



**Figure 1.** Predicted and actual journey mode (to school) by distance

Replication of the modelling over a large number of individuals enables population-level distributions of trip behaviour to be built up. The resulting time activity data can then be intersected with information on environmental hazards, such as air pollution or accident risks, to derive estimates of risks to health and well-being. Since the model is probabilistic in nature it cannot be said to model actual behaviour or exposures, except where detailed input information is provided. The time

activities generated represent scenarios, representing specific population groups and activity patterns. This method provides a powerful tool both for research and policy with diverse applications. They include studies of population dynamics, accessibility, environmental exposures and health risks, transport and land use planning. One of the main applications of the model is the analysis of potential impacts of different policies or interventions and scenario modelling.

**6. Acknowledgements**

The journey to school project was funded by ESRC Small Grant RES-000-22-2023

**References**

Aylin, P., Bottle, A., Wakefield, J., Jarup, L., and Elliott P. (2001). Proximity to coke works and hospital admissions for respiratory and cardiovascular disease in England and Wales. Thorax 56:228–233.

Ball W., LeFevre S, Jarup L., Beale L., 2008, Comparison of Different Methods for Spatial Analysis of Cancer Data in Utah, Environmental Health Perspectives, Volume 116, Number 8

Beale L., Field K., Picton P., Mattews H., (2006) Mapping for wheelchair users: Route navigation in urban spaces, *Cartographic Journal*, 43(1), pp 68-81

Briggs D. (2005). The role of GIS: coping with space (and time) in air pollution exposure assessment. *Journal of Toxicology and Environmental Health*, Part A, 68:1243–1261, 2005

Elliott, P., Briggs, D. J., Morris, S., de Hoogh, C., Hurt, C., Kold Jensen, T., Maitland, I., Richardson, S., Wakefield, J., and Jarup, L. (2001). Risk of adverse birth outcomes in populations living near landfill sites. B. Med. J. 323:363–368.

English, P., Neutra, R., Scalf, R., Sullivan, M., Waller, L., and Zhu L. (1999). Examining associations between childhood asthma and traffic flow using a geographic information system. Environ. Health Perspect. 107:761–767.

Cronberg A, Christophersen O, Pickering O, Tipping S, (2007) National Travel Survey 2006, Technical Report, Prepared for the Department for Transport.

Jarup L. (2004). Health and environment information systems for exposure and disease mapping, and risk assessment. *Environmental Health Perspectives* 112 (9):995-997.

Walker, M, Whyatt, J.D., Pooley, C, Davies, G, Coulton, P and Bamford, W, 2009. Talk, Technologies and Teenagers: Understanding the School Journey using a Mixed-Methods Approach. *Children's Geographies* 7 (2) 107-122.

**Biography**

*Linda Beale is a Research Fellow in GI Science and Health at Imperial College London with over ten year's research experience in GIS. Her research interests are primarily in the development of methodologies for spatial analysis, modelling and programming bespoke GIS applications for epidemiological study.*

*Duncan Whyatt is a Senior Lecturer in GIS at Lancaster University with research interests in air pollution at a variety of spatial and temporal scales.*

*Frederico Fabbri has recently been employed as a Research Assistant at Imperial College London having just graduated with a first class degree in Computing and GIS.*

*Gemma Davies is the Spatial Information Support Officer at Lancaster University.*

*David Briggs has a chair in Environmental Sciences and Health at Imperial College London and has led the GIS group in the UK Small Area Health Statistics Unit for a number of years.*

# Integrating Real-time Bus-Tracking with Pedestrian Navigation in a Journey Planning System

Bashir Shalaik, Ricky Jacob, Adam C. Winstanley

Department of Computer Science, National University of Ireland Maynooth, Ireland
Tel. +353 1 708 3853
bsalaik@cs.nuim.ie, rjacob@cs.nuim.ie, adam.winstanley@nuim.ie

KEYWORDS: Real-time vehicle tracking, Automated vehicle location, Pedestrian navigation, Journey planners,

## 1. Introduction

Automated Vehicle Location (AVL) systems provide real-time location information for emergency response, delivery services and freight transport. The advent of AVL systems has meant both public and private bus operators can implement systems to provide real-time passenger information, analyse their service performance and also to evaluate the quality of their operations. Traffic congestion, intersection delays, weather and operational conditions are some of the factors that make it difficult to predict the accurate bus arrival time in a real-time environment. In a joint project between NUI Maynooth and Blackpool Transport, a dynamic web application was developed to display and update vehicle locations (bustracking.co.uk) (Winstanley et al. 2009) and to provide predictive bus arrival times at stops.

A journey by bus is usually part of a longer door-to-door itinerary, usually involving walking before, after or between bus segments. The passenger is really interested in door-to-door journey times when making decisions about time of departure and which bus to catch. Therefore journey planners that combine the pedestrian and bus journeys are required and indeed several such systems exist, such as Transport Direct (2009), Traveline Midlands (2009), Google transit (2009). However these systems are mainly designed to plan journeys in advance and so base their decisions on the fixed bus timetable. For last-minute planning, and also for updating journey plans as-you-go, real-time bus locations and short-term predictions of bus arrival times at stops can be used to give more reliable journey times taking into account delays due to congestion, diversions and other factors. This paper describes an experimental system that combines bus tracking and pedestrian navigation.

## 2. Bus Tracking, Pedestrian Navigation and Journey Planning

Recent advances of geo-positioning hardware, computer software and mobile communications have combined to offer new opportunities for improved public transport services. Today many public transport agencies are using vehicle tracking to provide travellers with detailed, reliable, high-quality, real- time travel information. Mostly these use the Global Positioning System (GPS) and wireless communication systems (for example, radio data systems or GSM/GPRS) for communicating their vehicle location information and other details to a central server (figure 1). By tracking their bus fleet in real-time, operators can monitor schedule adherence and service efficiency, give better operational support and provide users with real-time service information. There are several bespoke systems commercially available that do this. These systems can also build up an archive of data that can be analysed and mined for information to show the behaviour of the transport system over time, indicating recurrent problems such as vehicle bunching and delays due to congestion. In addition, to qualify for public subsidies, operators must report Quality-of-Service (QoS) metrics to regulatory authorities. These are usually calculated manually but the existence of a full archive of data gives the potential for automation.

In a joint project between NUI Maynooth and Blackpool Transport, a dynamic web application has been developed to display and update vehicle locations (bustracking.co.uk) (Winstanley et al. 2009), to provide predictive bus arrival times at stops and also to automatically calculate QoS metrics. The system uses off-the-shelf GPS/GPRS integrated units programmed to transmit locations at regular intervals while the vehicle is in motion. The data is stored on a server and can be displayed through a standard web browser. The system is implemented using web technologies such as JavaScript, MySql, XML, PHP and Ajax.

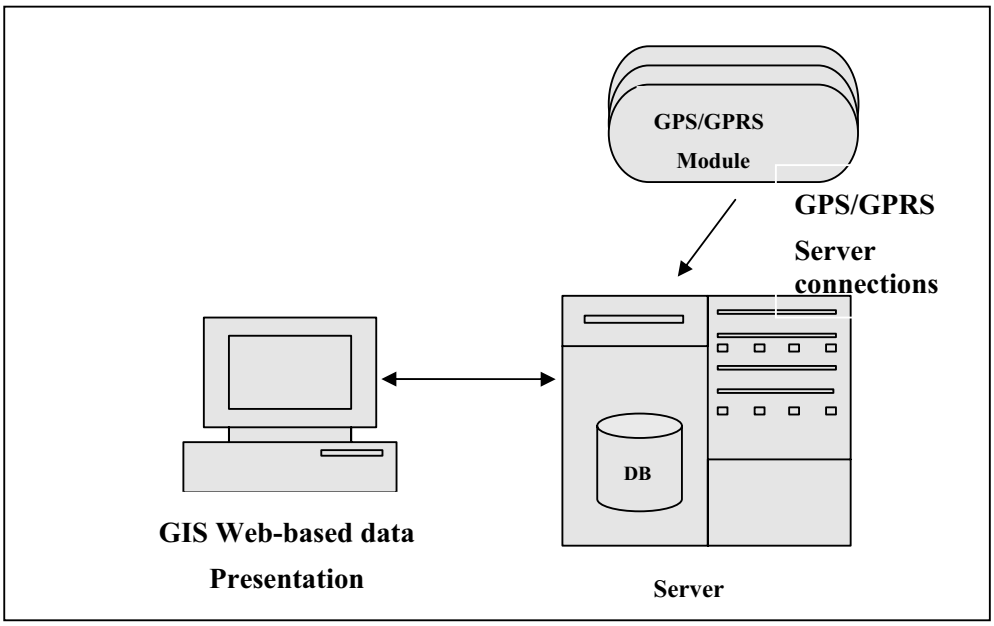



**Figure 1.** Data collection system

The GPS/GPRS units installed on buses provide location, direction and speed information. This is used to calculate how closely vehicles are following scheduled routes. Vehicle locations are displayed on a web base map (Google, MS Bing or OpenStreetMap) using icons automatically colour-coded to show how closely they are following their schedules. The same real-time data is used to generate displays showing predicted bus arrival times at stops.

Navigation dominates the greater part of the mobile mapping market. A recent article (Arrington 2009) claims that navigation takes over 70% of the $2 billion worldwide mapping market. This has resulted in diverse user -oriented systems and applications that are focused on specific application domains (urban pedestrian, car, freight, and hiking). Most of the big mobile mapping providers (Google, Nokia, Microsoft, Yahoo and Navteq) include pedestrian navigation as a part of their service. Pedestrians have different requirements from vehicle navigation. They are not restricted by rules of the road; they are free to follow a greater variety of paths; they can cross open spaces; their low speed means that routing instructions need to be more detailed; they tend to have personal preferences (for example, avoiding busy roads or preferring not to cross open spaces in bad weather) that affect the route they take; and they move seamlessly in and out of buildings. At NUI Maynooth, a campus pedestrian navigation system is being developed (Ciepluch 2009) that takes these factors into account, creating user profiles for customised route finding and communicating the route to the user in non-cartographic interactions such as audio or haptic feedback.

No major city is without a comprehensive on-line journey planning system that usually not only returns routes but also uses the timetable to give a detailed timed itinerary. However, a bus trip is

usually part of a longer door-to-door journey with walking segments before, after and/or between bus rides. A few applications such as Transport Direct, TravelLineMidlands and Google transit provide the user with the functionality to compute complete journeys. However, these applications are mainly designed for planning journeys in advance using the bus schedule as given in the fixed timetable. As such they cannot take into account delays and cancellations or provide re-routing instructions on the fly.

### 3. Integrating Real-time Bus-Tracking with Pedestrian Navigation

An experimental system integrating bus tracking and pedestrian navigation into the same application has been created. The system is implemented using Cloudmade and Web Map Lite which uses OpenStreetMap data in their map tiles and provides an API which has various services like routing and geo-coding. The system provides door-to-door routing and timing information for the specified journey and is designed to be used during the journey to provide up-to-date best-journey information based on the real-time location of buses.

Given the current location of a vehicle we have assessed three different prediction models for estimating arrival times at bus-stops using:

- historical data to determine typical travel time over that segment of the route at the same time-of-day and day-of-week,
- a multiple regression model and
- a Kalman filter model based on recent bus locations.

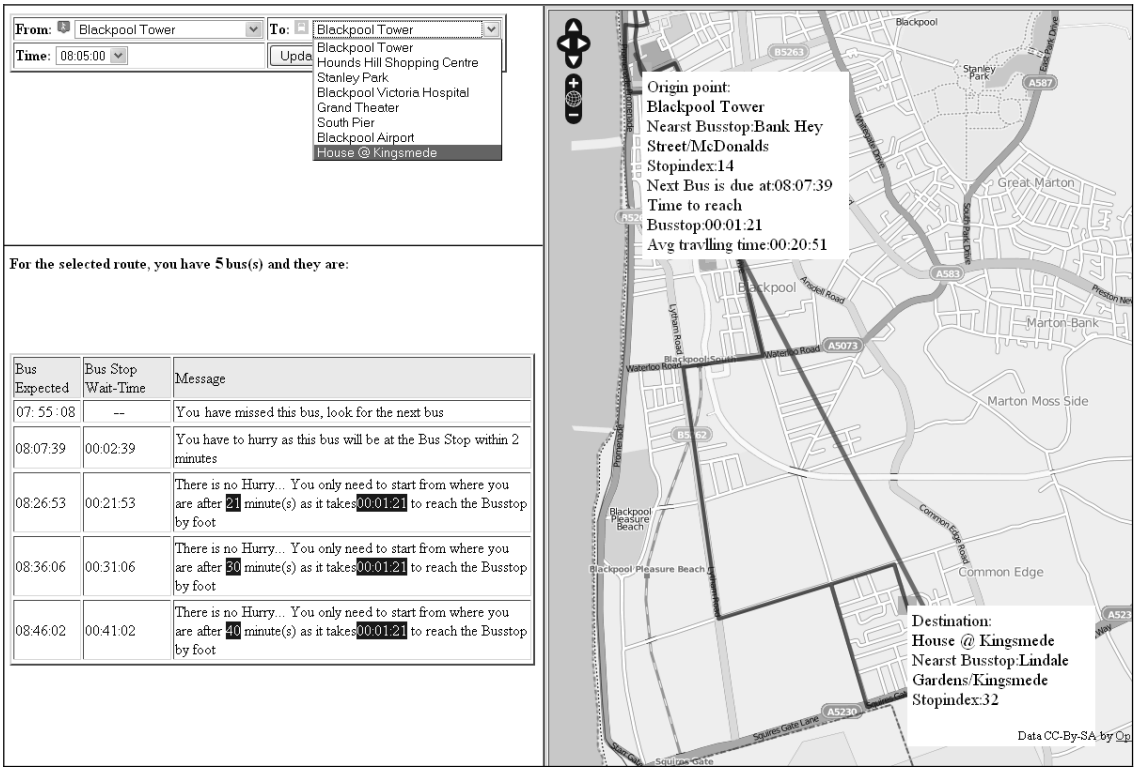
In order to evaluate the performance of these predictions models, the Mean Absolute Percentage Error (MAPE) was used to measure the closeness between predicted and observed values. MAPE (Ren and Glasure 2009) represents the average percentage difference between the observed value (actual arrival time) and the predicted value (table 1). The historical data model has been shown to produce the smallest MAPE.

In the experimental system, the user inputs their start point (or current location) and destination and the system finds the nearest bus stop that he needs to go to take the bus to get to his destination. The system calculates the time taken to take the shortest path to the bus stop and, after calculating the next predicted bus arrivals at the stop, suggests departure times to synchronise with bus departures and the resulting total travel time.

| Model | MAPE |
|---|---|
| Historical data Model | 13 |
| MLR Model | 29 |
| Kalman Filter Model | 20 |

**Table 1**: MAPE values of Prediction Models

The system uses OpenStreetMap and Cloudmade to interact with the user but could use any of the common web mapping systems (Google, Bing, Yahoo). Figure 2 shows the web interface of the current system presenting a typical journey specification with several trip options.




| Bus Expected | Bus Stop Wait-Time | Message |
|---|---|---|
| 07: 55 : 08 | -- | You have missed this bus, look for the next bus |
| 08:07:39 | 00:02:39 | You have to hurry as this bus will be at the Bus Stop within 2 minutes |
| 08:26:53 | 00:21:53 | There is no Hurry... You only need to start from where you are after 21 minute(s) as it takes 00:01:21 to reach the Busstop by foot |
| 08:36:06 | 00:31:06 | There is no Hurry... You only need to start from where you are after 30 minute(s) as it takes 00:01:21 to reach the Busstop by foot |
| 08:46:02 | 00:41:02 | There is no Hurry... You only need to start from where you are after 40 minute(s) as it takes 00:01:21 to reach the Busstop by foot |

**Figure 2.** Bus tracking and pedestrian navigation in an integrated journey planning system

## 4. Conclusion

This system brings together the functionality of an AVL system and pedestrian navigation with bus arrival time prediction to provide the user a real-time door-to-door journey planning system. This system uses a historical data model for bus journey time prediction. We intend to make the web page more interactive and give the user options to select destination points on the map rather than from a list. We also plan to extend the system to make it work with a GPS-enabled mobile phone where the start point will be the current position of the user and the user only specifies the destination point. In addition we plan to incorporate the full pedestrian navigation options currently included in our campus navigation system.

## 5. Acknowledgements

Thanks are due to Blackpool Transport Ltd for facilitating this project and particularly to Oliver Howarth. This project was partly supported by the StratAG (Strategic Research in Advanced Geotechnologies) project, funded under the SFI Strategic Research Cluster Programme (07/SRC/I1168). One author is supported by a PhD studentship from the Libyan Ministry of Education.

**References**

Arrington, Michael (2009): How Cloudmade Will Deal With Google Navigation Monster http://www.techcrunch.com/2009/10/30/how-cloudmade-will-deal-with-google-navigation-monster/- last accessed November 2009.

Ciepluch B, Jacob R, Mooney P and Winstanley A (2009) ACM SIGSPATIAL GIS 2009 PhD showcases -Using OpenStreetMap to deliver location-based environmental information in Ireland *SIGSPATIAL Special* 1(3) 17-22.

Ramakrishna Y, Ramakrishna P, Lakshmanan V and Sivanandan R (2006) Bus travel time prediction using GPS Data, *Map India* 2006.

Ren, L. and Glasure, Y. (2009) Applicability of the revised Mean Absolute Percentage Errors (MAPE) approach to some popular normal and non-normal independent time series, *International Advances in Economic Research*, 15 (4), 409-21, 2009.

Winstanley, A, Shalaik, B, Zheng, J and Burke, R: Visualizing public transport quality of service, *GISRUK 2009*, Durham, April 2009.

Weakliam, J., M. Bertolotto, and D. Wilson (2005) Implicit interaction profiling for recommending spatial content , Geographic Information Systems, *13th annual ACM international workshop on Geographic Information Systems*, Bremen, Germany, 285-294.

**Biography**

*Bashir Shalaik is a postgraduate research student at National University of Ireland, Maynooth having obtained an MSc degree in Computer Science at the same institute. He is interested in Intelligent Public Transportation and GPS-based vehicle tracking systems.*

*Ricky Jacob is a postgraduate research student at National University of Ireland, Maynooth. He completed his MSc at University of Madras, India. He is interested in novel guidance systems such as audio and haptic feedback for pedestrian navigation.*

*Adam Winstanley is the Head of Computer Science at National University of Ireland Maynooth and a senior research associate with the National Centre for Geocomputation. His research interests lie in several fields, including location based systems, intelligent transport systems; graphics recognition and electric vehicle control systems.*

.

# A Network Data Model for traffic simulation

José R. G. Alvarado[1], José L. F. Pereira[2], Rosaldo J. F. Rossetti[1]

Artificial Intelligence and Computer Science Laboratory (LIACC)
[1]Department of Informatics Engineering
[2]Department of Electronics and Computer Engineering (DEEC)
Faculty of Engineering, University of Porto (FEUP)
Rua Dr. Roberto Frias, S/N, 4200-465 Porto, Portugal
(+351) 220402900
ramon_tt@hotmail.com, ee06201@fe.up.pt, rossetti@fe.up.pt

KEYWORDS: Network Data Model, traffic simulation, GIS-T, Transportation, LRS

## 1. Introduction

Network modeling and analysis have become a real practical problem in the continuing growing cities. Facing daily traffic and its impact on the environment and society behavior are nowadays issues calling for suitable solutions. In order to provide these solutions we rely on *Geographical Information Systems* (GIS) and more precisely the branch that concentrates in *Transportation* (GIS-T). GIS-T must be capable to store, manage, process, retrieve and display network and traffic information gathered from multiple resources as GPS devices, shape files, and so on, as more accurately and closer to reality as possible, yielding to the necessity of an adequate data model to fulfill these requirements.

One of the most popular approaches in GIS-T modeling is the *arc-node* view where nodes are generally represented by points denoting road intersections and arcs are line segments serving as road segments. In respect to this model Goodchild (2000) identifies some problems:

- The indispensability of lane level road segments definitions for modern highway simulation demands and applications.
- Intersections issues such as, roundabouts and intersections that does not cross physically.
- No multimodal routing available.
- Redundancy, since street names must be repeated for each *arc.*

These problems affect the exactitude and certainty in which data is modeled thus generating inaccurate simulations and results. In pursuance of real life applicable solutions it is indispensable to address these issues.
In this paper we introduce a model that tries to give solution to the problems mentioned above expanding the arc-node model adding features to generate a richer structured model. The model was implemented as part of the MAS-T$^2$er Lab (Rossetti et al. 2008) traffic simulator in conjunction with the Network Editor (Pereira et al. 2009) in which some figures where designed.

## 2. The arc-node view and the linear referencing system

The network model is not a whole new concept, as it appears in the pioneer research in GIS modeling. Goodchild (1992) establishes two stages for its definition. In the first stage the network is created with nodes and arcs. In the second stage, further information such as gas stations, traffic lights, etc., is embedded into the network. This data may be expressed as the tuple $T = < l, o, z_1, z_2, \ldots, z_n >$ where $l$

stands for the *link* which data is connected to, whether a node or an arc, *o* is the offset from the origin of the link, and *z* variables represent the properties of the object being modeled. As an example, we can mention a speed limit sign. In this case, the link would be a road, the offset could be the milepost at which the object is placed, and a property could be an integer value for the speed limit. This process for constructing a network and locate objects is also known as *Linear Referencing System* (LRS) (Blazek 2004).

It is clear that for our traffic simulation purposes this arc-node model is incomplete, and must be extended to provide us with all the functionalities expected. However, the locating mechanism is still valid for us, since the precision used for static objects as traffic sings, traffics lights, hospitals, schools, etc., is somehow acceptable. In the next part we will just consider the data model extension as the main focus of our study.

## 3. The proposed model

In Figure 1 the Network data model designed is depicted.

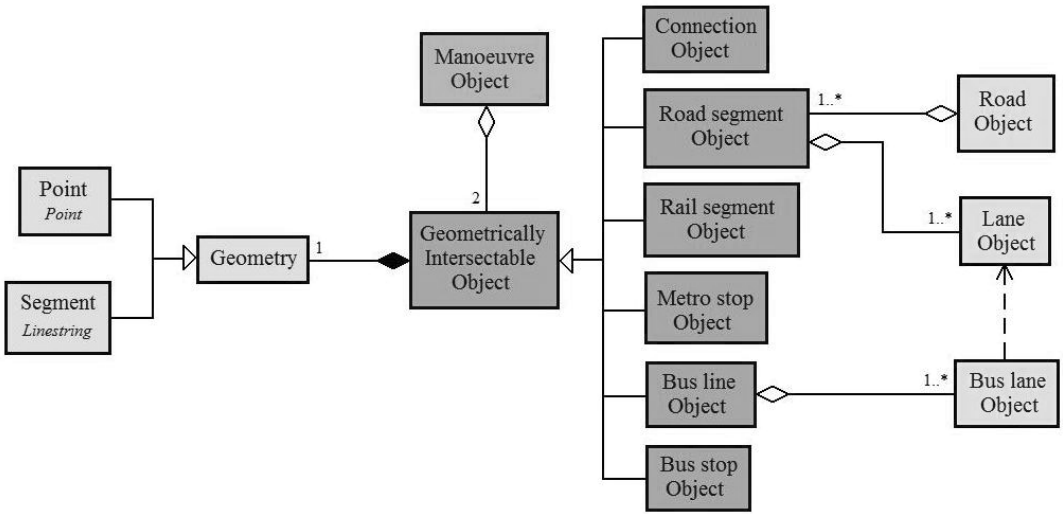



**Figure 1.** Structure of the Network data model proposed.

We shall describe all the elements of the model and how these interact for solving regular LRS problems. Firstly we have the *Geometry*. That can be either a *Point* or a *Segment*. The Point is represented by the OGC spatial object *Point* (OGC Reference model 2009). The Segment is denoted by the OGC spatial object *Linestring* (OGC Reference model 2009). One important reason to choose Linestring for Segments is because segments are expected to describe roads, and those frequently take irregular curve shapes as shown in Figure 2.

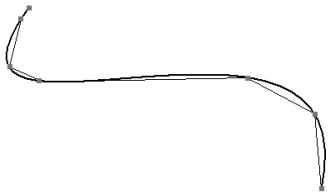


**Figure 2.** Irregular curve discretized to line segments.

Second comes what we called the *Geometrically Intersectable Object* (GIO). The GIO is composed of exactly one Geometry, hence it is accepted as parameter in the spatiotemporal function *ST_Intersects(Geometry,Geometry)* (OGC Reference model 2009); from here the term intersectable is suggested. A GIO can take the following forms:

- *Connection object:* A simple segment with no type that allows connections between other GIOS.
- *Road segment object:* A segment that represents a part of a road.
- *Rail segment object:* A segment that stands for a part of a railway, e.g. in a metro line.
- *Metro stop object:* A point denoting a metro station. It lies at end of begin of a rail segment.
- *Bus line object:* A set of *Bus Lanes*.
- *Bus stop object:* A point that match with a point in a Bus line.

Third, we introduce the concept of *Manoeuvre*. A Manoeuvre allows a route between GIOs, which means it is possible to travel from one GIO to another verifying that those geometrically intersect using the functions ST_Intersects maintaining in this way, the data integrity. In this place it can be realized the usefulness of Connection objects. In a roundabout for example, the incoming Road segments do not cross each other, but there is a Manoeuvre from that permits the connections as illustrated in Figure 3.

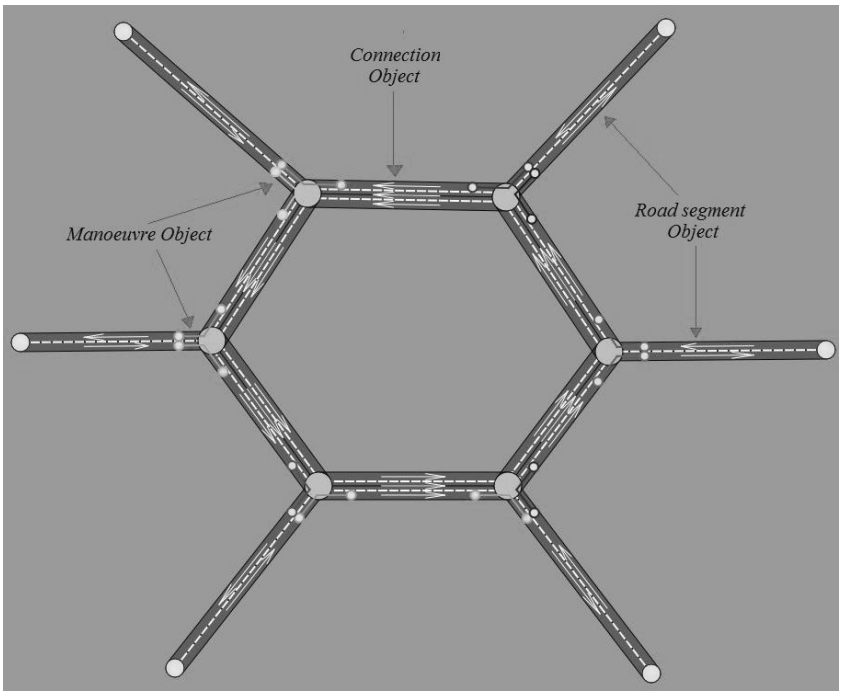
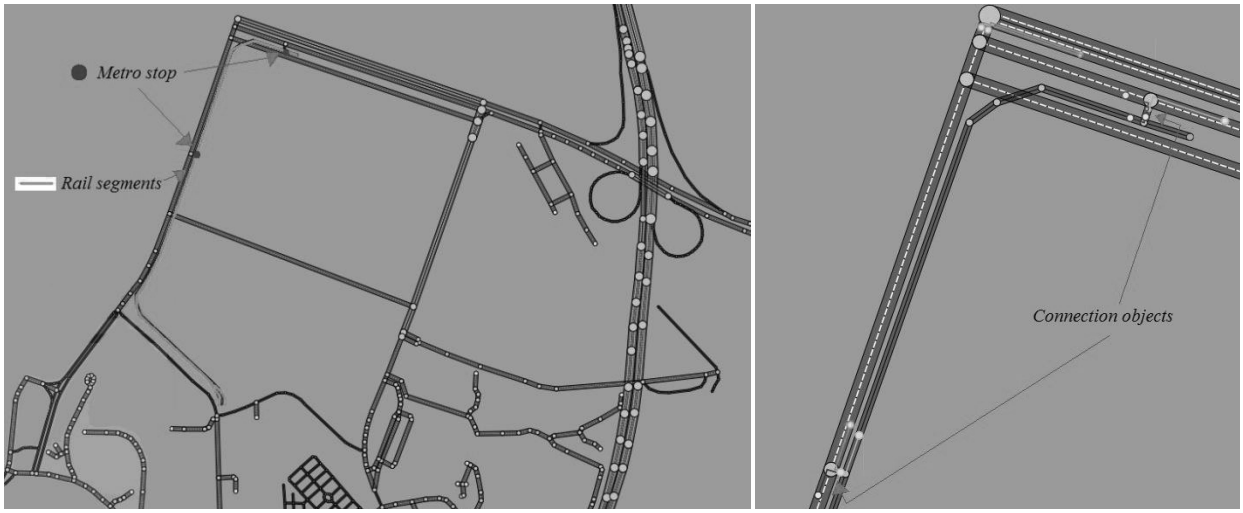



**Figure 3.** Roundabout showing how Manoeuvres connects different GIOS

Of course, this is in case the Connection objects have no particular name; otherwise those will be contemplated as Road segments. With the specified Manoeuvres we can find a route from network elements that do not geometrically intersect, but there exist a successive chain of GIO elements that enables a path between them. Manoeuvres are thus the solution for most intersection dilemmas.
However Connection objects would rather be used in small pieces, such as shown in Figure 4. In this case the Connection objects enable a route from a Road segment to a Metro stop. This is applicable if a pedestrian simulation is desired. Consequently, we can state that Connection object solve the multimodal routing problem since we can establish connections with every element in the network.




(a) (b)

**Figure 4.** Sao Joao hospital zone in Porto, Portugal.
(a) In green, the rail segment on which the metro city travels. In blue, Sao Joao and I.p.o. metro stations.
(b) The small connection objects that enable a path from a road segment to a metro stop are illustrated.

To continue, we have the *Road object.* This object is intended to solve the redundancy problem by constructing a whole *Street* taking a set of Road segments. In that manner the street name appears just once as Road object property and not in every Road segment.
Finally, we explain the *Lane and Bus Lane objects*. The lane object provides the granularity required for lane level data modeling and processing. A lane can get different properties such as: *Bicycle lane, Pedestrian Lane, etc.* We already include the dependent class *Bus Lane* to be able to create complete *Bus lines*. It is important to notice that every lane is also a GIO, for that reason Manoeuvres are also built at lane level.

Now, with this detail of information, we can export our data to other traffic simulators such as *SUMO* (Simulation of Urban Mobility) (Krajzewicz et al. 2002), as exposed in Figure 5.

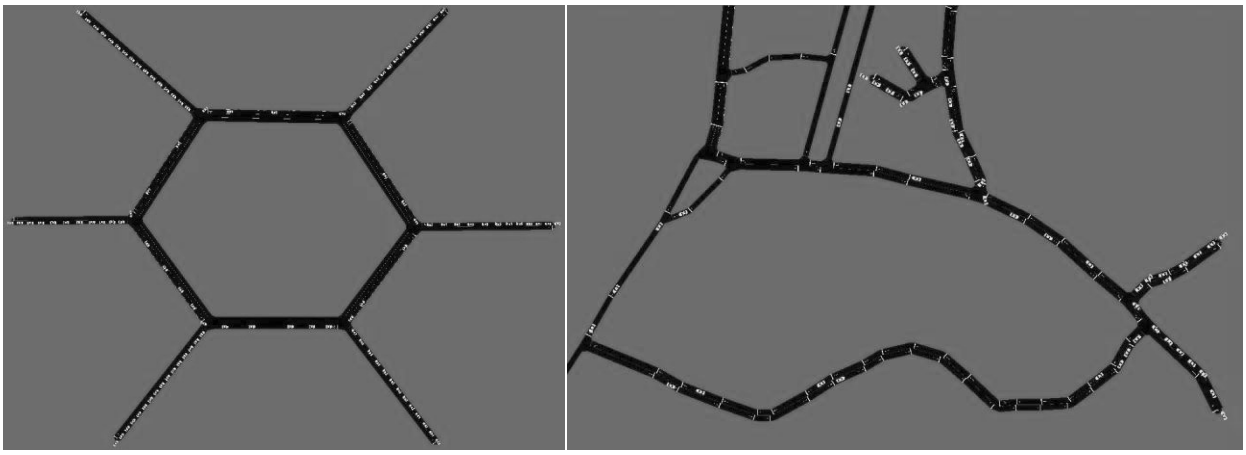


**Figure 5.** Examples of traffic simulations in SUMO.

## 4. Related Work

The ArcGIS Network Analyst team presented a technical paper (ESRI 2005) where they address some issues when exporting data for analysis such as duplicated records, street connectivity, and turn tables (manoeuvres). The described data model attempts to eradicate these problems from the very beginning in data insertion, in a manner that the exported data complies to be consistent and no further processing is required.

The model is well-suited and was implemented in PostgreSQL DBMS and PostGIS spatial data extension. We obtained very favorable results using these open source software tools in respect of modeling. In which corresponds to Routing functionality and Location Based Services these are supported by PostLBS. From here should be noted that modeling transportation network data focus in storing and managing the information, thus evaluating and analyzing could be complemented by external tool such as PostLBS.

## 5. Conclusions

Traffic simulation requires a very deep data abstraction able to yield very complex data models for complex problems. However as ITs grow, it is feasible to handle enormous amounts of information. The proposed model takes advantage of these capabilities and creates a more detailed data representation, maintaining consistency, integrity and accuracy of the collected data. Naturally this does not represent a definitive model; it can still be extended to increase the multimodal routing features available. Further work can also emphasize on adding lanes types, such as bicycle and pedestrian as mentioned above. The complexity levels of the proposed model can also be discussed. The idea in Data Network Modeling is to go as far and accurate as possible applying exhaustive methods for processing information, with the very last intention of create very rich data model that help to understand the underlying problem of traffic city network in society.

## 6. Acknowledgements

We acknowledge the outstanding assistance from Professors A. Coelho and P. Tavares giving us important insights into GIS and traffic network modeling.

## References

Blazek, R., (2005), 'Introducing the Linear Reference System in GRASS', *International Journal of Geoinformatics*, **1(3)**, pp. 95-100.

ESRI (2005, September), '*Preparing Street Data for Use with the Network Dataset'*, Technical paper J-9484. Retrieved February 28, 2010, from http://downloads2.esri.com/support/whitepapers/other_/J9484_Street_Data_w_Network_Dataset.pdf

Goodchild, M. F., (1992) 'Geographical data modeling', *Computers & Geosciences*, **18(4),** pp. 401-408.

Goodchild, M. F., (2000) 'GIS and transportation: status and challenges', *GeoInformatica*, **4(2)**, pp. 127-139.

Krajzewicz, D., Hertkorn, G., Rössel, C., Wagner, P., (2002); 'SUMO (Simulation of Urban MObility); An open-source traffic simulation', *Proceedings of the 4th Middle East Symposium on Simulation and Modelling (MESM2002)*, SCS European Publishing House, pp. 183-187.

*OGC Reference model*. Retrieved November 20, 2009, from http://www.opengeospatial.org/standards/orm

Pereira, J. L. F., Rossetti, R. J. F., Oliveira, E. C., (2009) 'Towards a Cooperative Traffic Network Editor', *CDVE2009: The 6th International conference on cooperative design, visualization and engineering*.

Rossetti, R. J. F., Ferreira, P., Braga, R., Oliveira, E. C., (2008) 'Towards an artificial traffic control system', *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems*, Beijing, China, October 12-15, pp.14-19.

## Biography

José R. G. Alvarado is a student in the Instituto Tecnológico Durango (ITD) in México, currently working as external assistant in the Faculty of Engineering of Porto University (FEUP). His areas of interest and research are robotics, chess engines and geoinformatics.

José L. F. Pereira is a student in the Faculty of Engineering of Porto University (FEUP). His areas of interest are open source and robotics. He is currently developing the Traffic Network Editor.

Rosaldo J. F. Rossetti is assintant professor in the Department of Informatics Engineering at FEUP and also a Research Fellow Artificial Intelligence and Computer Science Laboratory, LIACC. He is currently involved in the MASTER Lab Project, which is about developing agent-based technologies for ITS research.

# Positional accuracy and spatial linkage for a new global health dataset: GPS clusters in the Egyptian Demographic and Health Survey

Shawky Mansour,[a,b] David Martin,[a] Jim Wright [a]

[a]School of Geography, University of Southampton, Southampton SO17 1BJ, UK,
Tel +44(0) 238059 5000 Fax +44 (0)23 8059 3131

[b]Assistant lecturer at the Department of Geography and GIS, Alexandria University, Alexandria, Egypt

KEYWORDS: GIS linkage, EDHS, GPS cluster, spatial error

## 1. Introduction

This paper describes work undertaken as one component of a wider project to analyse the effects of access to safe drinking water on health in Egypt. Demographic and Health Survey (DHS) and census data are being used to examine the relationship between drinking water and health. The DHS is a global survey, launched in 1984 by the United States Agency for International Development (USAID) covering 84 developing countries (MEASURE DHS, 2009a). In the DHS, Global Positioning System (GPS) technology has been used to georeference households. The survey, which covers demographic and health topics, thus offers an important spatial dataset in developing countries. Montana and Spencer (2004) describe GPS data collection and use in the DHS: surveys are conducted with multiple households within local sampling units and these results are collectively referenced by a single 'GPS cluster' location. The DHS includes questions about health and water supply which are complementary to information about water supply and broader socioeconomic characteristics captured by the Egyptian census. There are good grounds for expecting variations in the quality of community-level water supply to be reflected in these datasets.

There is a growing literature on DHS applications using GIS and spatial analysis techniques, particularly in child mortality and disease (Uthman 2008, Wirth et al. 2006 and Gemperli et al.2004). However, there has been little coverage in the literature of linkage between DHS clusters and administrative boundary data. We were only able to identify one such study (Pande et al.2008) which linked DHS cluster points to other spatial datasets in an investigation of diarrhoea prevalence across Benin. This study neither explains in detail how DHS clusters were linked to administrative boundaries nor any problems identified during the point-in- polygon linkage process.

The utility of the DHS for our study would thus be considerably enhanced by linkage to census data, which provides coverage of the entire population. However, these datasets cannot be directly linked using existing established data tables, as might be possible in a UK setting GIS provide powerful tools for undertaking spatial linkage between map layers and in this case a route for the association of DHS results with census data. However, an assessment of the positional accuracy of the DHS and its implications for GIS-based linkage to census data is essential. Crosetto et al. (2000) describe positional error as the variation between measured and true phenomena in the real world. The spatial error that occurs in the input data of any GIS operation can propagate through to the outputs (Heuvelink 1998). We here concentrate on how to assess and accommodate spatial uncertainty in the HS so as to facilitate an important linkage between the DHS and census.

## 2. Method

### 2.1 Egyptian census geography

Egypt is divided into four main types of region: Upper or Valley Governorates, Lower or Delta Governorates, Urban Governorates, and Frontier or Desert Governorates. These contain twenty eight governorates plus Luxor Supreme Council City. The sub-national administrative boundaries consist of two further levels; *kism* or *markaz* and *Shyakha*. Kism (urban) and Markaz (both urban and rural) are the first level lower than governorate; Shyakhas are the smallest units in the Egyptian Census.

### 2.2 Data sources

The paper uses data from both the 2005 Egyptian DHS and 2006 national census. MEASURE DHS is the official organization undertaking the EDHS while the Egyptian Central Agency for Public Mobilisation and Statistics (CAPMAS) is responsible for censuses, surveys, and other public statistics. There are 360 Markazes and Kisms and 5705 Shyakhas in the 2006 Egyptian census and 1298 GPS clusters in the 2005 EDHS, covering all Egyptian governorates.

### 2.3 Assessment of spatial error

One objective of this study is to associate specific census and DHS variables and so understand potential impacts of drinking water availability on health in Egypt. For this purpose, the GPS cluster points and census polygons were overlaid in ArcGIS initially to examine the level of agreement between comparable DHS and census variables. The GPS clusters are not identified by individual kism or shyakha names so we are reliant on spatial linkage to associate data between the two sources. When overlaying Egyptian census polygon boundaries with GPS clusters, we found 10 cluster points that fell entirely outside Egyptian boundaries and 53 other points for which the governorate label in the DHS did not match the corresponding CAPMAS map layer. Therefore, at least 63 cluster points are either mislabelled or lie in incorrect positions (Figure1).

There is significant evidence of global clustering in GPS clusters with incorrect governorate labels. Exploring the pattern of GPS cluster errors further, the majority of misallocated points are located very close to governorate boundaries. We therefore need to understand more about the pattern of these errors in order to determine whether they can be corrected or whether they would introduce unacceptable levels of misallocation to each level of the census geography. DHS documentation indicates that in some datasets, a geographic error is randomly added to the cluster locations of up to 2 km for urban areas and 5 km for rural areas (MEASURE DHS, 2009b) in order to protect confidentiality although this is not specifically indicated for the 2005 Egyptian DHS. We therefore used logistic regression to model the effect of distance on governorate labelling errors and obtain the probability of each cluster point being assigned to an incorrect polygon. We also investigated whether the error pattern differed between rural and urban GPS clusters.

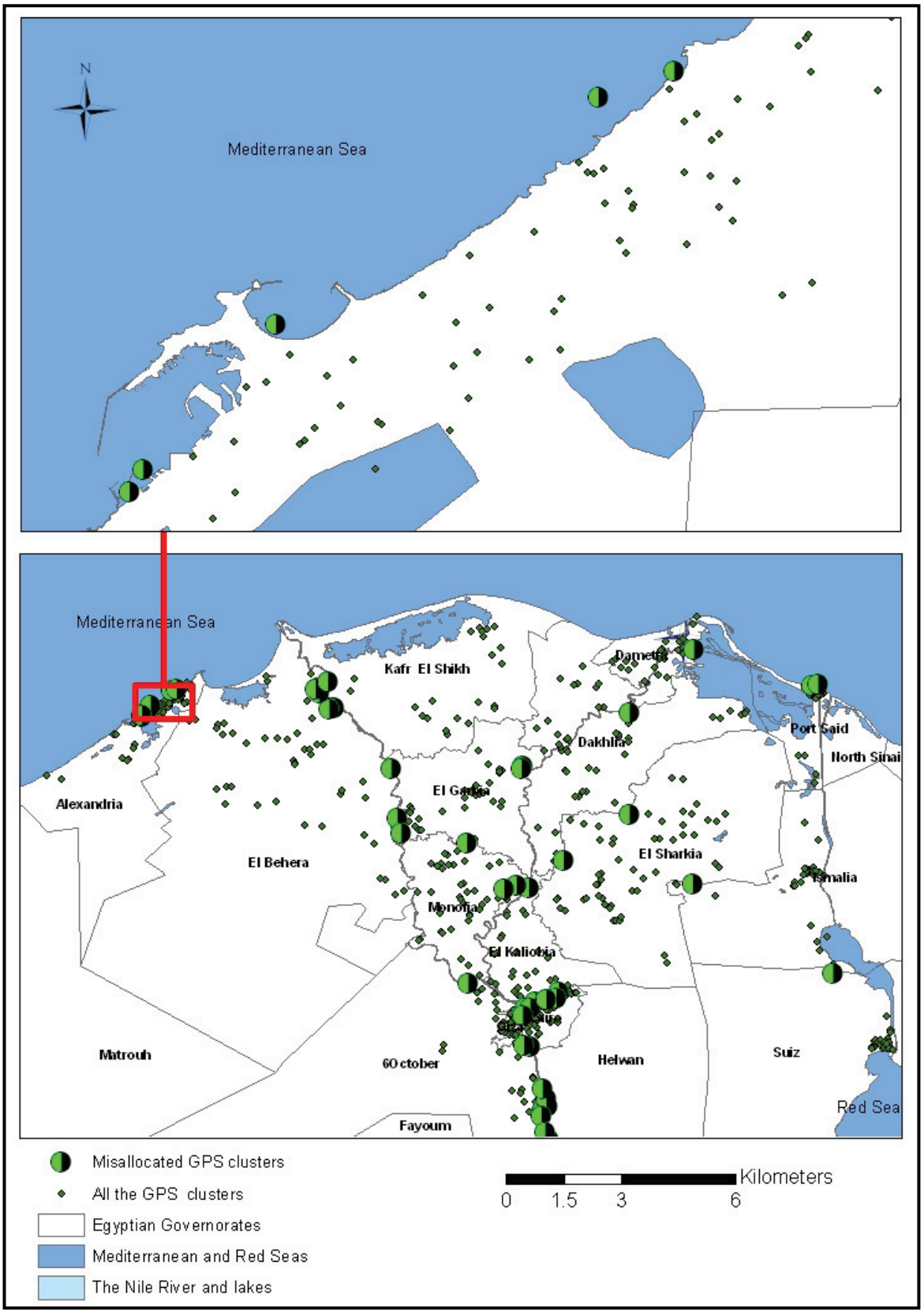



**Figure 1.** Spatial distribution of GPS clusters of EDHS with and without positional error in Delta Governorates and Greater Cairo

## 3. Results

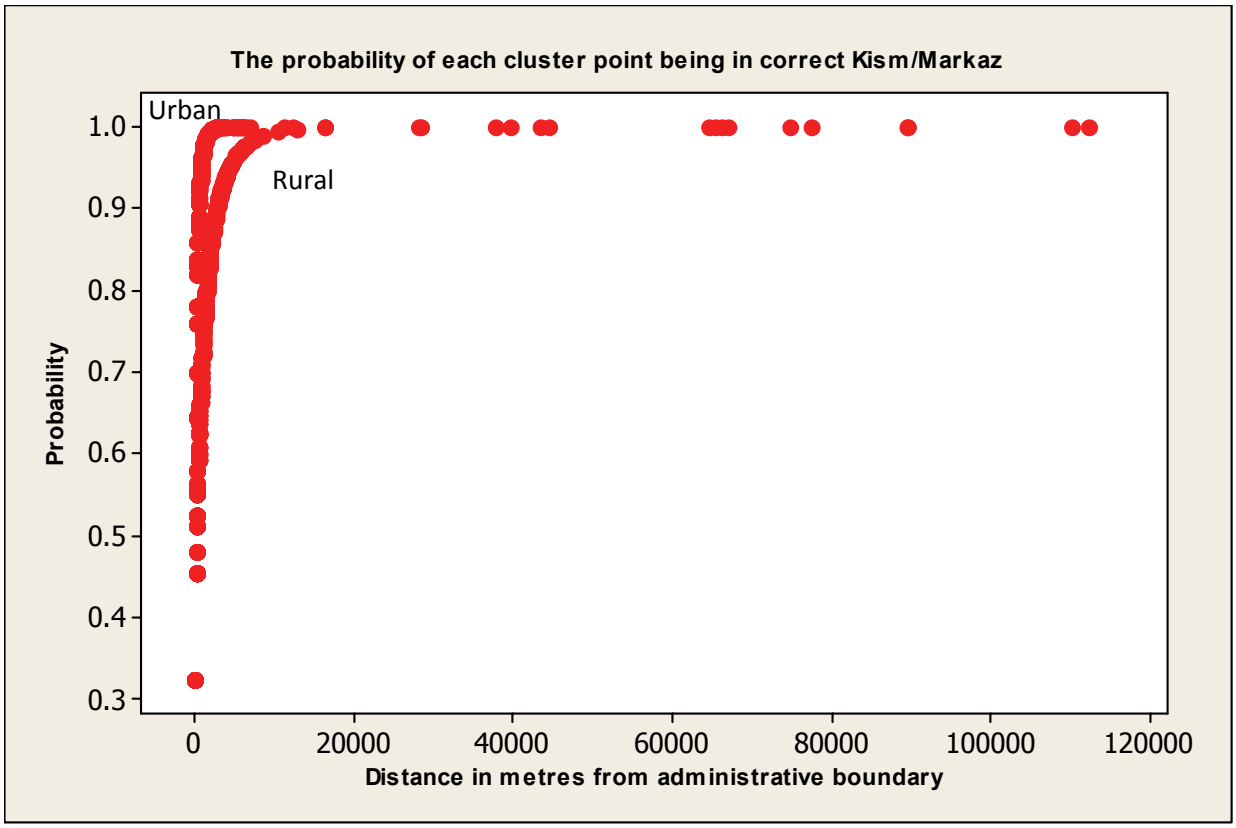
A series of logistic regression models was undertaken using different predictors (GPS survey methodology used, rural vs urban clusters, and distance to governorate boundary) and transformations to improve the strength of the model. The final model (Table 1), which best explained the observed errors, examines the dependent variable (misallocated cluster points) using two independent variables; square root of distance of all GPS cluster points to governorate boundary and an interaction variable created by multiplying this distance variable by a dichotomous rural-urban variable (where each cluster point is coded one if it is urban and zero if not). Thus, the distance of GPS cluster points from administrative boundaries affects the probability of each cluster being in its correct location. Figure 2 shows how the probability of cluster errors differs from urban to rural. This difference is likely to be caused by the deliberate scrambling of GPS cluster locations which is greater in rural areas than urban. The coefficients of this model can be used to predict the probability of spatial error based on location of GPS cluster points from sub-provincial boundaries (Kism/Markaz). By applying these coefficients to the lower level administrative boundaries, 7.87 % of the GPS clusters showed a low probability value (less than 0.5) of being in their correct position. Hence, a considerable number of DHS GPS clusters would fall within an incorrect Kism or Markaz. *Shyakha* linkage process was also undertaken. However, at this scale there would be unacceptable levels of uncertainty and consequently, unreliable matching between Shyakhas and GPS clusters. Thus, the Kism/Markaz census geography appears to be the best geographic level for undertaking spatial linkage.

| *Variables* | *Coef.* | *Std.Err* | *Z value* | *95% conf.* | **P> Z** |
|---|---|---|---|---|---|
| **√(Governorate distance)** | -0.056 | -8.37 | -8.37 | -0.069 | 0.000 |
| **Urban*squared distances** | -0.078 | 0.015 | -5.08 | -0.108 | 0.000 |
| **Constant** | 0.740 | 0.299 | 2.48 | 0.154 | 0.013 |

**Table 1.** The output of regression model

**Log likelihood** = -154.438 **Prob > chi2** = 0.000 **Pseudo R**2 = 0.39 **LR chi2 (1)** = 195.22

**Figure 2.** The probability of being in correct Kism/Markaz according to the regression model presented in Table 1




## 4. Conclusion and future work

The addition of GPS clusters to the DHS offers increased analytical potential in GIS and global health research. However, our analysis suggests administrative labels and spatial locations of GPS clusters cannot be used uncritically and error may affect spatial analysis of DHS datasets. This is most likely due to deliberate modification of cluster locations noted earlier, intended to protect the confidentiality of participant households in the DHS, although this procedure is not documented for the EDHS. A strong recommendation to the MEASURE DHS organization would be to adapt the data modification algorithm to avoid gross misallocation between administrative areas. In combining DHS data from GPS clusters with census data for Kizms/Markazes to understand safe water access, any linkage process will need to take account of the likely impact of positional uncertainty. All researchers working with previously undocumented GPS-based datasets should consider investigation and modelling of spatial error before undertaking GIS linkage with administrative boundaries.

## 5. Acknowledgements

We would like to thank Egyptian Government and Department of Geography and GIS at Alexandria University for funding. We would also like to thank MEASURE DHS in the United States and Central Agency for Mobilisation and Statistics (CAPMAS) in Egypt for provision of datasets.

## 6. References

CROSETTO, M., TARANTOLA, S. & SALTELLI, A. (2000) Sensitivity and uncertainty analysis in spatial modelling based on GIS. Agriculture, Ecosystem and Environment, 81, 71-79.

GEMPERLI, A., VOUNATSOU, P., KLEINSCHMIDT, I., BAGAYOKO, M., LENGELER, C. & SMITH, T. (2004) Spatial patterns of infant mortality in Mali: the effect of malaria endemicity. American journal of epidemiology, 159, 64.

HEUVELINK, G. B. M. (1998) Error propagation in environmental modelling with GIS, London, CRC Press.

MEASURE DHS (2008a) DHS History. http://www.measuredhs.com/aboutdhs/history.cfm accessed on 25/12/ 2008.

MEASURE DHS (2008b) Geographic Information Systems, Methodology -Collecting Geographic Data http://www.measuredhs.com/aboutsurveys/gis/methodology.cfm accessed on 10th of January 2009.

MONTANA, L. & SPENCER, J. (2004) Incorporating geographic information into MEASURE surveys: a field guide to GPS data collection. Macro International Publication,http://www.measuredhs.com/basicdoc/gps/DHS_GPS_Manual.pdf accessed on 24/06/2009.

PANDE, S., KEYZER, M. A., AROUNA, A. & SONNEVELD, B. (2008) Addressing diarrhea prevalence in the West African Middle Belt: social and geographic dimensions in a case study for Benin. International Journal of Health Geographics, 7, 17.

UTHMAN, O. A. & KONGNYUY, E. J. (2008) A multilevel analysis of effect of neighbourhood and individual wealth status on sexual behaviour among women: evidence from Nigeria 2003 Demographic and Health Survey. BMC International Health and Human Rights, 8, 9.

WIRTH, M. E., BALK, D., DELAMONICA, E., STOREYGARD, A., SACKS, E. & MINUJIN, A. (2006) Setting the stage for equity-sensitive monitoring of the maternal and child health Millennium Development Goals. Bulletin of the World Health Organization, 84, 519-527.

## 7. Biography

*Shawky Mansour is a second year PhD student at the School of Geography, University of Southampton. He holds a BSc in Geography and MA in Political Geography from the Department of Geography, Helwan University, Egypt and MSc in GIS from University of Leeds. His PhD research focuses on GIS and public health particularly investigation of spatial relationships between drinking water and population health outcomes.*

# Versioning Administrative Geographic Data on the Semantic Web

Alex Lohfink[i], Duncan McPhee[ii]

University of Glamorgan, Wales Institute of Social and Economic Research, Data and Methods (WISERD), Pontypridd,CF37 1DL
Tel. +44443 482950
(alohfink, dmcphee)@glam.ac.uk

KEYWORDS: administrative geography, semantic web, versioning, RDF, linked data

## 1. Introduction

The Resource Description Framework (W3C 2004) has become an important language in the representation of distributed data on the semantic web. It has been successful in representing relationships between web resources at the data level, as opposed to the presentation level, enabling websites to publish machine-readable information about relationships between distributed resources, rather than relying on relational database-driven web pages that express relationships within queries.

RDF uses Uniform Resource Identifiers, or URIs, to identify resources, their properties and property values, on the web. These values can be represented by nodes and arcs of a graph. So, for example, the graph shown in Figure 1 could be used to represent the statements "there is civil parish identified by http://data.ordnancesurvey.co.uk/id/7000000000000005, which is called Chelmsley Wood, and has the census code 00ct006."

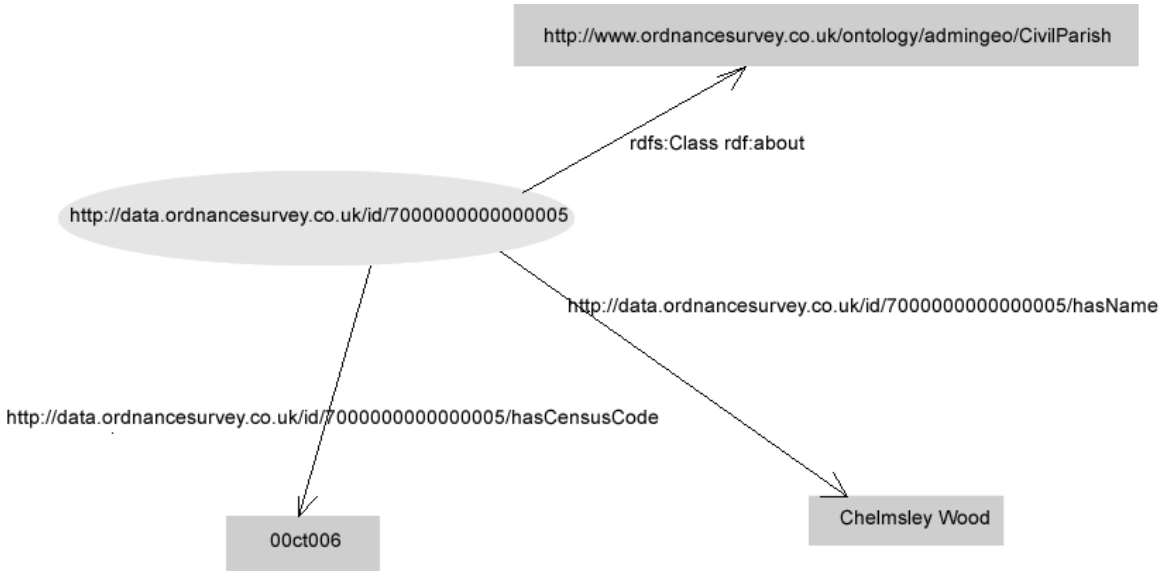



*Figure 1- RDF graph describing Chelmsley Wood*

At the heart of the RDF data model is the RDF triple. This is a simple structure based on three parts: subject: predicate: object, or alternatively, resource: property: value. An example of a triple from Figure 1 would be:

subject: Civil Parish (URI http://data.ordnancesurvey.co.uk/id/7000000000000005), predicate: census code (URI http://data.ordnancesurvey.co.uk/id/7000000000000005/hasCensusCode), object: literal (00ct006).

This triple represents the statement "There is a civil parish called Chelmsley Wood that has the census code 00ct006". Statements can thus be represented and linked to form a directed graph.

An issue with RDF data is that resources are subject, as with all data, to evolution through change, and this can lead to linked resources being either removed or outdated. In this paper we look at some possible methods that could be employed to version-enable RDF administrative geography data.

## 2. The need for versioning in administrative geographic RDF data

Versioning for RDF can be viewed from two perspectives: web ontology versioning (classes) and instance versioning (triples). Instance versioning can be further divided into model-based and statement-based (instance) versioning. Model-based versioning applies to a group of triples that form part of a logical unit. Statement-based versioning applies to individual statements (triples). It is within the field of instance versioning that this paper is concerned. Specifically, we aim to address a versioning issue that has arisen in Ordnance Survey[1] Administrative Geography data. Here, the relationship between administrative units in the UK is described by RDF data. However, there are many changes that have occurred to administrative units represented in this data over the years, and undoubtedly, many more will occur. At present, there is no way to represent different versions of administrative units in the data. Further, different users of the administrative geography datasets could link to different versions of administrative units represented in whichever version of the data they are accessing, meaning that inconsistencies will be apparent between different RDF datasets. It would therefore be beneficial to provide a versioning mechanism that would allow linkage to a default version of an administrative unit, but allow access to alternative versions if specified. At present, to our knowledge, there is no proposed system or model that provides this.

### 2.1 Previous work

There have been some attempts at introducing version mechanisms to RDF graphs, mainly centred on ontology versioning and determining differences between graphs. The SemVersion (Volkel 2005) model focuses on managing change in ontologies where users can suggest different classes to include in the ontology. SemVersion can manage such changes and reconcile them into a new version of the ontology. SemVersion employs model-based versioning.

Delta (Berners-Lee and Connolly 2001) is a system designed to identify differences between RDF graphs, and uses functions to compute these differences. Differences between graphs are produced in the form of a *delta* which represents the changes only. This means that a *delta* produced from a knowledge base can be applied to a subset of this knowledge base and update it, with accurate results.

Another version model described by (Ludwig, Kuster et al. 2008) uses an extension to the Topic Maps data model (ISO 2008) to potentially implement versions in RDF. Topic Maps represent topics (or subjects), attributes, and associations as an entity-relationship model. The (Ludwig, Kuster et al. 2008) model uses a structure called the *VersionInfo Object*, or *VIO*, to record start and end dates for a specific version of a topic map object. This model is stated as being applicable to RDF triples by grouping triples into logical units and linking them to a *VIO*.

Both Delta and SemVersion are aimed at managing change to web ontologies or specific RDF graphs rather than addressing the need to be able to reference different versions of the same statements or classes within the same RDF dataset. (Ludwig, Kuster et al. 2008) attempts to provide a method for achieving this, by linking logical units of triples to VIOs. In this case, the suggestion is that the VIO would contain start and end dates relevant to the statements in the referenced logical unit, in effect extending the graph to incorporate version objects. However, no implementation is specified, and it is not clear how alternatives would be handled. It also organises VIOs according to a sequence, the organisation of which is not specified.

---

[1] This research is sponsored by Ordnance Survey

## 3. Possible mechanisms for versioning RDF data

### 3.1 RDF containers

RDF containers provide the capability to represent collections of triples as single entities, and link to them forming new triples. This means that it may be possible to model relationships between multiple versions of data by defining appropriate RDF container classes. Of particular interest here is the *rdf:Alt* container. The *rdf:Alt* element is used to describe a list of alternative values, the first value in the list being the default value.

### 3.2 Inferencing

RDF Schema (RDFS, (W3C 2004)) allows inferencing based on defined properties such as *subClassOf* and *subPropertyOf* . At the simplest level, this provides a mechanism to create a version hierarchy based on inheritance, where new versions of an item are defined using the *subClassOf* (or 'is_a') relationship. Inferencing allows RDF to infer from the *subClassOf* relationship that the resource is a member of the superclass. For example, In Ordnance Survey Administrative Geography data, a Civil Parish is defined as a *subClassOf* a Civil Administrative Area. It can thus be inferred that the Civil Parish of Chelmsley Wood is also a Civil Administrative Area. This feature provides type propagation, and could be used to define a version hierarchy of RDFS classes.

It is also possible with some RDF implementation environments to define inferencing rules. This kind of inferencing goes beyond the scope of the RDFS inferencing capabilities, and allows the definition of specific, text-based rules by which implicit relationships can be inferred. This could allow version-specific rules to enhance a version hierarchy, such as *version_of*, *derivative_of*, *alternative_to* based on criteria derived from the differences between versions of an administrative unit.

### 3.3 Named graphs

Named graphs allow groups of triples to be identified as belonging to a specified named graph within a larger RDF graph. This is achieved by tagging the triples with an identifier that specifies the named graph to which it is associated, in effect making the triples "quads". This means that a group of triples could be coupled together as a named version graph. The RDF query language, SPARQL (W3C 2008), has a *FROM NAMED* clause which can query named graphs.

## Conclusions

In this report we have given a brief background to RDF, and discussed some of the possible methods which may be adopted for the purposes of versioning RDF at the instance (statement) level in administrative geography data. Previous work on versioning in this area has been largely centred on web ontologies and graph differences, but we have suggested here several methods that show the potential to implement versioning of RDF at the instance level. It is our intention next to implement a versioning system by producing a prototype web application using one or more of these methods.

## References

Berners-Lee, T. and D. Connolly. (2001). "Delta: an ontology for the distribution of differences between RDF graphs." Retrieved 3/11/2009, from http://www.w3.org/DesignIssues/Diff.

ISO. (2008). "Information Technology - Topic Maps - Part 2: Data Model " Retrieved 3/11/2009, from http://www.isotopicmaps.org/sam/.

Ludwig, C., M. W. Kuster, et al. (2008). Versioning in Distributed Semantic Registries. iiWAS2008**:** 493-499.

Volkel, M. (2005). "SemVersion – Versioning RDF and Ontologies." Retrieved 3/11/2009, from http://semversion.ontoware.org/kwebd233a.pdf.

W3C. (2004). "RDF Vocabulary Description Language 1.0: RDF Schema." Retrieved 3/11/2009, from http://www.w3.org/TR/rdf-schema/.

W3C. (2004). "RDF/XML Syntax Specification (Revised)." Retrieved 3/11/2009, from http://www.w3.org/TR/rdf-syntax-grammar/.

W3C. (2008). "SPARQL Query Language for RDF." Retrieved 3/11/2009, from http://www.w3.org/TR/rdf-sparql-query/.

---

[i] Alex Lohfink is a lecturer at the University of Glamorgan, and gained a PhD there in December 2008. His current research interests are the semantic web and spatio-temporal databases.

[ii] Duncan McPhee is a senior lecturer at the University of Glamorgan. His research interests are in computer-based learning, databases, data mining, and the semantic web.